BIAS REDUCTION AND GOODNESS-OF-FIT TESTS IN CONDITIONAL LOGISTIC REGRESSION MODELS

A Dissertation

by

XIUZHEN SUN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2010

Major Subject: Statistics

BIAS REDUCTION AND GOODNESS-OF-FIT TESTS IN CONDITIONAL LOGISTIC REGRESSION MODELS

A Dissertation

by

XIUZHEN SUN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Suojin Wang
	Samiran Sinha
Committee Members,	P. Fred Dahm
	Jianxin Zhou
Head of Department,	Simon J. Sheather

August 2010

Major Subject: Statistics

ABSTRACT

Bias Reduction and Goodness-of-Fit Tests in Conditional

Logistic Regression Models.

(August 2010)

Xiuzhen Sun,

B.S., Shandong Normal University;

M.S., Southern Methodist University;

M.S., Texas A&M University

Co–Chairs of Advisory Committee: Dr. Suojin Wang
Dr. Samiran Sinha

This dissertation consists of three projects in matched case-control studies. In the first project, we employ a general bias preventive approach developed by Firth (1993) to handle the bias of an estimator of the log-odds ratio parameter in conditional logistic regression by solving a modified score equation. The resultant estimator not only reduces bias but also can prevent producing infinite value. Furthermore, we propose a method to calculate the standard error of the resultant estimator. A closed form expression for the estimator of the log-odds ratio parameter is derived in the case of a dichotomous exposure variable. Finite sample properties of the estimator are investigated via a simulation study. Finally, we apply the method to analyze a matched case-control data from a low-birth-weight study.

In the second project of this dissertation, we propose a score typed test for checking adequacy of a functional form of a covariate of interest in matched case-control studies by using penalized regression splines to approximate an unknown function. The asymptotic distribution of the test statistics under the null model is a linear combination of several

chi-square random variables. We also derive the asymptotic distribution of the test statistic when the alternative model holds. Through a simulation study we assess and compare the finite sample properties of the proposed test with that of Arbogast and Lin (2004). To illustrate the usefulness of the method, we apply the proposed test to a matched case-control data constructed from the breast cancer data of the SEER study.

Usually a logistic model is needed to associate the risk of the disease with the covariates of interests. However, this logistic model may not be appropriate in some instances. In the last project , we adopt idea to matched case-control studies and derive an information matrix based test for testing overall model adequacy and investigate the properties against the cumulative residual based test in Arbogast and Lin (2004) via a simulation study. The proposed method is less time consuming and has comparative power for small parameters. It is suitable to explore the overall model fitting.

I dedicate this work to my parents and my family.

ACKNOWLEDGMENTS

I would like to take this opportunity to express my sincere gratitude to my two advisors, Professors Suojin Wang and Samiran Sinha, for their immense help at every stage of my research. I remain grateful for their constant encouragement and helpful suggestions in my studies.

My appreciation extends to my other committee members, Professors P. Fred Dahm and Jianxin Zhou, for their time and thoughtful comments on my research and dissertation. I also would like to thank all the other professors in my department for instructing me throughout my graduate studies. I appreciate the kindness of Professors Michael Longnecker and Simon J. Sheather for writing letters in my job search during their busy work schedule. I give my special thanks to Professor Michael Longnecker for his wonderful suggestions and guidance when I first started to teach at Texas A&M University.

I am greatly indebted to my many family friends for encouraging and supporting me spiritually throughout difficult times. I learn from them how to have a peaceful life.

My deep love and gratitude go to my parents for their unconditional love. They devote their whole life to me and teach me how to live better:

"Forgiving others is to release yourself",

"Giving makes you happier than taking".

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION

1.1 Study Designs in Epidemiology

Epidemiology is the study of the distribution and determinations of diseases in human populations. One of goals of Epidemiology is to assess the relationship between disease and potential risk factors of interest, which sometimes are called exposures, or covariates. Primarily, there are two types of design to collect data. If we apply "treatments" to the subjects, for example, give patients certain doses of medicine to see how it affects the disease, then it is called an experimental design. If we just observe what are happening and record the information from a group of subjects without imposing "treatments" to the subjects, this type of design is called an observational study.

Observational studies are commonly used in epidemiology. In observational studies, generally data are collected through *cross-sectional, prospective cohort*, or *case-control* study. In a cross-sectional study, the disease outcome and potential risk factors are observed at a given point in time. So it is also called an epidemiological survey. It provides a snapshot of the frequency and characteristics of a disease in a population on a particular point in time. Cross-sectional study is useful in showing association between different variables and can provide early clues to etiology. Usually it is faster and costs less. However, in some cases it may be difficult to determine which are the effects and which are the causes. For example, if we find that people with cancer are more likely to have heavy drinking problems, but we cannot tell whether heavy drinking problems cause cancer or people with cancer are more likely to drink a lot as a comfort. Furthermore, if the disease is rare, we may not observe

---

This dissertation follows the style of *Biometrics*.

sufficient number of diseased subjects in the sampled data.

In a prospective cohort study, a group of study subjects (cohort) with heterogeneous exposures, or two or more groups defined by certain exposure status is obtained, and then the incidence of the outcome is recorded during the follow-up study period. If the disease is rare, we may end up with only a few subjects with disease, which will lead to a less powerful statistical test to test the hypothesis of interest. Furthermore, there exists a low likelihood that the population may take a long time to develop the disease. For example, the study of death from lung cancer could involve 20 to 40 years, potentially longer than the careers of many epidemiologists. Thus the study generally will cost a huge amount of money due to the time consumed and a large number of subjects involved in the study. However, a prospective cohort study can provide stronger evidence of causality and provide potential risk factors with less bias due to errors of recall or measurement.

Case-control studies are types of epidemiological designs which compare individuals who have disease (cases) with a group of individuals without the disease (controls). When there is convincing evidence about the association between disease and potential exposures, the related sources of exposures are reallocated for studies. Case-control studies are particularly suited to investigate the risk factors for rare diseases and diseases that take a long time to manifest. Case-control studies require much smaller sample sizes than the usual prospective studies and can deal with multiple risk factors simultaneously. Generally, case-control studies are very informative. Once a population based disease is identified, we can describe the picture of the disease, for example, we can estimate the incident rate according to subjects' age, gender, location, etc.

The distinction between a case-control study and a prospective study lies in the sampling. In a case-control study we sample from among the diseased and nondiseased, whereas in a prospective study we sample from among those with the factors of interest and those without the factors.

1.2 Research Topics

The two main topics in the inferential analysis of a study are parameter estimation and hypothesis testing. Estimation involves the use of information from a sample to represent the measurement from the target population, while the hypothesis testing is a statistical method to discover whether the assertion about the population is believable based on the sample information. In the hypothesis testing, we first set up a null hypothesis denoted by H_0 against an alternative hypothesis denoted by H_1, then seek information from samples to check whether the data support the null hypothesis H_0 statistically. If the data is not compatible to H_0, then the test rejects H_0. Otherwise, it fails to reject H_0 which means there is no strong evidence to detect whether H_0 is true or false.

My dissertation researches on the point estimator and goodness-of-fit tests in case-control studies, specifically the work are done in the context of matched case-control settings.

1.3 An Example in Matched Case-Control Studies

It often happens that a case-control study involves some confounding variables. Confounding variables are extraneous factors that wholly or partially accounts for the observed effect of the risk factors on disease status. It can cause the association between disease and risk factors to appear or mask their true association. We can assess confounding by estimating the effect of the risk factors with and without allowing for confounding. Control of confounding can be achieved by stratified analysis. In stratified case-control studies, we compare the potential risk factors between the group of cases and the group of controls within homogeneous categories of the confounding variables.

Matched design is a special case of case-control studies where a case is matched with one or more controls within each stratum. Here is an example (see Table 1) that motivates

my research. It is about infants' low-birth-weight data from Hosmer and Lemeshow (1989), collected at the Baystate Medical Center, Springfield, Massachusetts, in 1986. Low birth

Table 1. Low-birth weight data.

STR	OBS	AGE	LOW	LWT	SMOKE	HT	UI	PTD
1	1	16	1	130	0	0	0	0
1	2	16	0	112	0	0	0	0
1	3	16	0	135	1	0	0	0
1	4	16	0	95	0	0	0	0
2	1	17	1	130	1	0	1	1
2	2	17	0	103	0	0	0	0
2	3	17	0	122	1	0	0	0
2	4	17	0	113	0	0	0	0
.	.	.	.	.	.	.	.	.
29	1	32	1	105	1	0	0	0
29	2	32	0	121	0	0	0	0
29	3	32	0	132	0	0	0	0
29	4	32	0	134	1	0	0	1

weight is one of the main concerns to the physicians. If the baby has low birth weight, generally the baby has high risk of death and also may suffer lifelong disabilities. So it is important to identify the potential risk factors that cause the low birth weight. Scientists believe that mother's behavior during pregnancy can play a major role for her baby's birth weight, thus the mothers' related information was measured. For example, LWT is the mother's weight at her last menstrual period, SMOKE denotes whether the mother smoked or not during her pregnancy, PTD is the status whether the mother has previous preterm delivery history, etc. For the detail of the data, one can refer to Hosmer and Lemeshow (1989).

An infant is defined as a case if its birth weight is below 2500 gms, otherwise the infant is a control. In this dataset mother's age is used as a matching variable to reduce the

potential effect due to the variation of age. For this dataset mother's age is between 16 and 32 years. Totally there are 29 strata.

In matched case-control studies, the probability of getting disease given the potential risk factors on a stratum is usually modeled by a logistic regression model with a common set of slope parameters and an intercept term which is a nuisance parameter that depends on the stratum. The common slope which are interpreted as log-odds-ratios are the parameters we are interested in. The log-odds ratio measures the degree of association between disease and potential risk factors. One of the study goals is to estimate the log-odd ratio for a mother to have a low-birth weight baby.

1.4 Bias and Maximum Likelihood Estimators

In statistics, an estimator denoted by $\widehat{\boldsymbol{\theta}}$ is a function of observed data that used to estimate the unknown population parameters denoted by $\boldsymbol{\theta}_0$. To estimate $\boldsymbol{\theta}_0$, we generally need to select a random sample from the target population, then calculate the point estimator $\widehat{\boldsymbol{\theta}}$. The value of $\widehat{\boldsymbol{\theta}}$ varies from sample to sample. Theoretically we need sample infinite times to form the distribution of the point estimator $\widehat{\boldsymbol{\theta}}$, then use the center $E(\widehat{\boldsymbol{\theta}})$ of distribution to estimate $\boldsymbol{\theta}_0$. The error of the estimator is defined as $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ and bias is defined as $b(\widehat{\boldsymbol{\theta}}) = E(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_0$, i.e., the expectation of the error. If $b(\widehat{\boldsymbol{\theta}}) = 0$, then $\widehat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}_0$. Otherwise it is biased.

In a logistic regression model, the log-odds ratio is obtained by maximizing the likelihood function. The Likelihood function is a function of parameters for statistical inference. If a sample $x_1, x_2, \cdots, x_n$ of n independent observations are drawn from probability density $f(\cdot|\boldsymbol{\theta})$, then the likelihood function is defined as

$$L(\boldsymbol{\theta}|x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})\,.$$

The point estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ called the maximum likelihood estimator (MLE) is obtained by maximizing log likelihood function

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \mathrm{argmax}_{\boldsymbol{\theta}} \log\left\{L(\boldsymbol{\theta}|x_1, x_2, \cdots, x_n)\right\} = \mathrm{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\left\{f(x_i|\boldsymbol{\theta})\right\} .$$

For a large sample size, the MLE has the following asymptotic properties:

- strong consistency: $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} \to \boldsymbol{\theta}_0$ as $n \to \infty$;
- asymptotic normality: the distribution of MLE $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ is convergent to a normal distribution with mean $\boldsymbol{\theta}_0$ and covariance matrix equal to the inverse of the Fisher information matrix.

The consistency and asymptotic normality property hold only under regularity conditions. Here are the conditions for 1-dimensional parameter which can be extended to multivariate cases (Serfling, 1980). For $\theta \in \Theta$, where Θ is an open interval,

1. $\log\{f(x|\theta)\}$ is three times continuously differentiable in $\theta \in \Theta$;
2. for each $\theta_0 \in \Theta$, there exists functions $g(x)$, $h(x)$ and $q(x)$ such that in some neighborhood of θ_0

   $$\left|\frac{\partial\left\{\log f(x|\theta)\right\}}{\partial\theta}\right| \le g(x), \quad \left|\frac{\partial^2\left\{\log f(x|\theta)\right\}}{\partial^2\theta}\right| \le h(x), \quad \left|\frac{\partial^3\left\{\log f(x|\theta)\right\}}{\partial^3\theta}\right| \le q(x)$$

   and

   $$\int g(x)dx < \infty, \quad \int h(x)dx < \infty, \quad E_\theta\left\{q(x)\right\} < \infty\,;$$

3. for each $\theta \in \Theta$,

   $$0 < E_\theta\left\{\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2\right\} < \infty\,.$$

The maximum likelihood estimator selects the parameter value which is mostly likely relative to the other values. The MLE is invariance under transformation, i.e., if $\phi(\boldsymbol{\theta})$ is any transformation of $\boldsymbol{\theta}$, then $\phi(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})$ is the MLE of $\phi(\boldsymbol{\theta})$.

However, when sample sizes are not large enough, the estimator could be significantly biased. One of the contributions of the dissertation is to deal with the bias problems in conditional logistic regression models.

1.5 Goodness-of-Fit Tests

The goodness-of-fit of a statistical model describes how well the model fits a sample of observations. Typically the discrepancy between the observed values and expected valued under the model is summarized to measure the goodness-of-fit. This can be achieved by testing the null hypothesis H_0 that a given random variable X follows a specified distribution $f(x; \boldsymbol{\theta})$.

In assessing whether a given distribution is suited to a dataset, we can use Kolmogorov-Smirnov test, Cramérvon-Mises criterion, and Anderson-Darling test based on the empirical distribution function. An attractive feature of these tests is that the distributions of the test statistics do not depend on the underlying cumulative distribution function which is being tested. However, in some cases these tests have low power. Another test statistic used very often is a chi-square test when the variance of the measurement error is known. The test statistic asymptotically follows a chi-square distribution with certain degrees of freedom for large sample sizes. The test rejects the null H_0 if the test statistic is larger than the critical value of chi-square distribution at specified significant level.

In a general model setting, likelihood ratio test is used to compare the fit of two models that one is nested within the other. The likelihood ratio Λ is the ratio of the likelihood function varying the parameters over two different sets in the numerator and denominator. Under H_0, when the sample size is large, $-2\log(\Lambda)$ converges to a chi-square distribution with degrees of freedom equal to the difference of dimensions of the null space and the alternative space.

Besides likelihood ratio test, there are other two commonly used methods of testing hypotheses concerning nested models. One is called a Wald test which compares the difference between maximum likelihood estimates of a group of parameters and their null values in relation to their variance. The other is a score test which uses the derivative of log likelihood evaluated at the null hypothesis. All these three methods are asymptotically equivalent. In this dissertation we construct a generalized score test to handle model fitting in conditional logistic regression models.

1.6 Organization of This Dissertation

Due to the presence of infinite dimensional nuisance parameters, matched case-control studies are distinct from the standard logistic regression analysis. The standard estimator of log-odds ratio parameter is obtained by maximizing the conditional likelihood function. However, for small to moderate sample sizes, the standard estimator is usually biased. For some data configuration, the estimate of the parameter could be infinite.

Furthermore, in epidemiological research often we assume that the covariates are associated with the disease risk through a linear-logistic model which means the logit of the disease probability is a linear function of the covariates. However, it may not be adequate to explain the effect of a continuous covariate on the disease risk.

My dissertation is designed to handle these issues in certain specified settings. This chapter has given a brief review of the research background. Chapter II is intended to be an expository introduction to some of the basic methods in case-control studies. Chapter III studies the methodology to deal with the bias problem in estimating log-odds ratio parameters. A general bias reduction approach developed by Firth (1993) is employed to reduce the bias of estimator of the log-odds ratio parameter in a matched case-control study by solving a modified score equation. In Chapter IV, we propose a generalized score

test to check whether a specific functional form of a covariate is adequate in the logistic regression model against a general unknown function form by applying a penalized spline function to approximate the unknown function such that the null model is a special case of the alternative. Furthermore, in order to test if the overall model is adequate to describe the disease risk, a generalized information matrix based method is developed for testing overall goodness-of-fit of the logistic model for matched case-control studies in Chapter V by following the idea of White (1982) that dealt with the effect of model misspecification on maximum likelihood estimators. Chapter VI offers concluding remarks and potential future research topics.

CHAPTER II

CASE-CONTROL STUDIES

2.1 Introduction

A case-control study is a retrospective design to establish an association between the presence of risk factors and the occurrence of a disease. It is useful for rare diseases or when the disease takes a very long time to become manifest. In case-control studies, the outcome is measured now and exposure is estimated from the past.

The present chapter is a brief review of the development of case control studies and methodology of analysis with different types of exposures, such as discrete or continuous variables. In the following we will use risk factors, exposures, or covariates alternatively to describe the presence of characteristics.

2.2 Tests for Homogeneity of Odds Ratio

The relative risk is defined as the ratio of the risk of a disease for those with risk factors to those without risk factors. If the relative risk is larger than 1, then the factors under investigation increase the risk; If the relative risk is less than 1, then factors reduce the risk. In epidemiological studies, we can evaluate the relative risk that individuals who have certain risk factors develop a specified disease, or we can measure the odds ratio. Odds is be the ratio of risk relative to the non-risk given the same risk factors, thus the odds ratio is the ratio of odds of the disease among the exposed subjects to the odds of the disease among the non-exposed subjects. An odds ratio above 1 implies the exposure to risk factor increases the odds of disease and a value of less than 1 means that the exposure reduces the odds of disease.

If Y is the binary disease indicator ($Y = 1$ for disease, $Y = 0$ for non-disease) and X

represents a dichotomous risk factor ($X = 1$ if risk factor presents, $X = 0$ if risk factor is absent), then the odds ratio is defined as

$$\begin{aligned}\theta &= \frac{\text{pr}(Y=1|X=1)/\text{pr}(Y=0|X=1)}{\text{pr}(Y=1|X=0)/\text{pr}(Y=0|X=0)} \\ &= \frac{\text{pr}(Y=1|X=1)\text{pr}(Y=0|X=0)}{\text{pr}(Y=0|X=1)\text{pr}(Y=1|X=0)}\,.\end{aligned}$$

Now let us consider a case-control study with a binary exposure variable X and disease status Y. Suppose there are n_1 cases and n_0 controls, and n_{11}, n_{10} are number of observations of factor present, and absent cases while n_{01}, and n_{00} are number of observations of factor present, and absent controls in the sample (see Table 2). Using the Bayes

Table 2. Case-control data on binary outcomes.

	Factors		
Disease Status	Factor Present	Factor Absent	Total
Case	n_{11}	n_{10}	n_1
Control	n_{01}	n_{00}	n_0
Total	$n_{.1}$	$n_{.0}$	n

rule (Bayes, 1763) $\text{pr}(A|B) = \text{pr}(B|A)\text{pr}(A)/\text{pr}(B)$,

$$\theta = \frac{\text{pr}(X=1|Y=1)\text{pr}(X=0|Y=0)}{\text{pr}(X=0|Y=1)\text{pr}(X=1|Y=0)}\,, \tag{2.1}$$

and thus it can be estimated by

$$\widehat{\theta} = \frac{n_{11}n_{00}}{n_{10}n_{01}}\,.$$

If the disease is rare, both $\text{pr}(Y=1|X=0)$ and $\text{pr}(Y=1|X=1)$ are close to 1, thus

$$\theta \approx \frac{\text{pr}(Y=1|X=1)}{\text{pr}(Y=1|X=0)}\,,$$

i.e., the odds ratio reduces to the relative risk.

Observe that $\theta = 1$ implies $\text{pr}(Y=1|X=1) = \text{pr}(Y=1|X=0)$, i.e., there is no

association between the disease and the risk factor.

For a large sample size, $\log(\widehat{\theta}) - \log(\theta)$ is asymptotically normally distributed with mean 0 and the variance $n_{10}^{-1} + n_{01}^{-1} + n_{11}^{-1} + n_{00}^{-1}$ (Woolf, 1955). The test of no association between disease and exposures is achieved by a chi-square test for large sample size. The test statistic with Yate's correction (Yates, 1934) is given by

$$\chi^2 = \frac{(n-1)\left(|n_{11}n_{00} - n_{10}n_{01}| - \frac{1}{2}n\right)^2}{n_1 n_0 n_{\cdot 1} n_{\cdot 0}}.$$

The test rejects null hypothesis that there is no association between disease and exposures if $\chi^2 > \chi^2_{1,\alpha}$ at α level of significance, where $\chi^2_{1,\alpha}$ is the $(1-\alpha)^{th}$ quantile of the χ^2_1 distribution.

For a small sample size, *Fishers exact test* procedure is used to test whether the odds ratio is equal to a specified value given that the row and column totals are fixed. Regardless of the data from case-control or a cohort study, the conditional probability on all the marginal totals remaining fixed is

$$\text{pr}(n_{11}|n_1, n_0, n_{\cdot 1}, n_{\cdot 0}; \theta) = \frac{\begin{pmatrix} n_1 \\ n_{11} \end{pmatrix} \begin{pmatrix} n_0 \\ n_{\cdot 1} - n_{11} \end{pmatrix} \theta^{n_{11}}}{\sum_u \begin{pmatrix} n_1 \\ u \end{pmatrix} \begin{pmatrix} n_0 \\ n_{\cdot 1} - u \end{pmatrix} \theta^u}.$$

Here $\begin{pmatrix} n \\ u \end{pmatrix}$ denotes the *binomial coefficient*, where

$$\begin{pmatrix} n \\ u \end{pmatrix} = \frac{n(n-1)(n-2)\cdots(n-u+1)}{u(u-1)(u-2)\cdots 1}.$$

Define the lower and upper tail probabilities respectively as follows

$$\begin{aligned} P_L &= \sum_{u \le n_{11}} \text{pr}(n_{11}|n_1, n_0, n_{.1}, n_{.0}; \theta_0)\,, \\ P_U &= \sum_{u \ge n_{11}} \text{pr}(n_{11}|n_1, n_0, n_{.1}, n_{.0}; \theta_0)\,. \end{aligned}$$

For a given level of significance α, the test rejects the hypothesis $H_0 : \theta = \theta_0$ in favor of $\theta < \theta_0$ if the lower probability $P_L < \alpha$. Similarly the test rejects the hypothesis H_0 in favor of $\theta > \theta_0$ if the upper probability $P_U < \alpha$.

2.3 Matched Case-Control Designs

If there is only one control matched with a case, it is called matched pairs. Suppose one has n_{10}, n_{11}, n_{01}, and n_{00} matched pairs under different combinations of X and Y, which are summarized in Table 3. The pairs in which both cases and controls are exposed to the

Table 3. Matched binary data on binary outcomes.

	Control		
Case($Y = 1$)	$X = 1$	$X = 0$	Total
$X = 1$	n_{11}	n_{10}	n_1
$X = 0$	n_{01}	n_{00}	n_0
Total	$n_{.1}$	$n_{.0}$	n

risk factor provide no information about the association between risk factor and disease. Similarly, the pairs in which neither case nor control are exposed to the risk factor provide no information. The statistical analysis depends on the *discordant* pairs, in which exactly one subject is exposed and other is not.

Let π be the conditional probability of observing a pair with an exposed case and

unexposed control given a discordant pair. Then

$$\pi = \frac{\text{pr}(X=1|Y=1)\text{pr}(X=0|Y=0)}{\text{pr}(X=1|Y=1)\text{pr}(X=0|Y=0)+\text{pr}(X=0|Y=1)\text{pr}(X=1|Y=0)} = \frac{\theta}{\theta+1},$$

where θ represents the odds ratio as defined in (2.1).

The conditional distribution of n_{10} given the total discordant pairs is a Binomial$(n_{10}+n_{01}, \pi)$. So the estimate of odds ratio θ of the disease and the risk factors is given by

$$\widehat{\theta} = \frac{n_{10}}{n_{01}}$$

with a standard error

$$s.e.(\theta) = \frac{n_{10}}{n_{01}}\sqrt{\frac{1}{n_{10}}+\frac{1}{n_{01}}}.$$

The test statistic to test $H_0 : \theta = 1$ is

$$\chi^2 = \frac{(|n_{01}-n_{10}|-1)^2}{n_{01}+n_{10}},$$

which is called *McNemar's test* (1947). Under the null hypothesis the test statistic asymptotically follows χ^2_1 distribution for large sample sizes. The test rejects H_0 if the value of χ^2 is larger than critical value $\chi^2_{1,\alpha}$ at level α, and concludes that the cases and controls differ in the presence of risk factors.

When M controls are matched to a case with dichotomous exposures, there are $2(M+1)$ possible outcomes depending on whether or not the case is exposed and number of exposed controls, therefore the conditional probability of the outcome that case is exposed is

$$\text{pr}(\text{case exposed}|m \text{ exposed among case and controls}) = \frac{m\theta}{m\theta+M-m+1}.$$

Assumption that there are totally m exposed in case and controls. Let $n_{1,m-1}$ be number of matched pairs that the case and exactly $m-1$ control are exposed, and $n_{0,m}$

be the number of matched pairs that exactly m control are exposed. Then $T_m = n_{1,m-1} + n_{0,m}$ represents the total number of matched sets with exactly m exposed. The conditional probability of the entire set of data is proportional to

$$\prod_{m=1}^{M} \left(\frac{m\theta}{m\theta + M - m + 1} \right)^{n_{1,m-1}} \left(\frac{M - m + 1}{m\theta + M - m + 1} \right)^{n_{0,m}} \tag{2.2}$$

with

$$\begin{aligned} E(n_{1,m-1}|T_m;\theta) &= \frac{m\theta T_m}{m\theta + M - m + 1}\,, \\ \text{Var}(n_{1,m-1}|T_m;\theta) &= \frac{m\theta T_m (M - m + 1)}{(m\theta + M - m + 1)^2}\,. \end{aligned}$$

The conditional maximum likelihood estimator (MLE) $\widehat{\theta}$ that maximize (2.2) is the solution of

$$\sum_{m=1}^{M} n_{1,m-1} = \sum_{m=1}^{M} \frac{m\theta T_m}{m\theta + M - m + 1}\,.$$

A simple robust Mantel and Haenszel common odds ratio estimator is given by

$$\widehat{\theta} = \frac{\sum_{m=1}^{M} (M - m + 1) n_{1,m-1}}{\sum_{m=1}^{M} m n_{0,m}}\,.$$

The corrected test statistic for testing $H_0 : \theta = 1$ derived by Miettinen (1970) and Pike and Morrow (1970) can be written as

$$\chi^2 = \frac{\left\{ \left| \sum_{m=1}^{M} (n_{1,m-1} - m T_m/(M+1)) \right| - \frac{1}{2} \right\}^2}{\sum_{m=1}^{M} T_m m (M - m + 1)/(M+1)^2}\,.$$

2.4 Logistic Regression in Case-Control Studies

So far we describe methods for having dichotomous expose variables. However, in epidemiological studies, disease is generally related to several different types of risk factors, such as categorical, continuous or mixture of both, thus a mathematical model is needed to

describe the relationship between the probability of disease and several risk factors. Usually a linear logistic model is used to deal with the risk of the disease in terms of risk factors. More detailed theory of logistic regression can be found in Cox (1970).

The general model of linear logistic model can be written as

$$\text{pr}(Y = 1|\boldsymbol{X}) = H(\alpha + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) \,,$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$, $Y = 1$ represents the individual who has disease and $Y = 0$ represents the individual free of the disease, and $\boldsymbol{X}$ is a vector of risk factors.

The odds for an individual to get disease given the same risk factors is

$$\psi = \frac{\text{pr}(Y = 1|\boldsymbol{X})}{\text{pr}(Y = 0|\boldsymbol{X})} = \frac{\text{pr}(Y = 1|\boldsymbol{X})}{1 - \text{pr}(Y = 1|\boldsymbol{X})} = \exp\left\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\right\} \,,$$

thus

$$\boldsymbol{\beta} = \log\left\{\frac{\text{pr}(Y = 1|\boldsymbol{X} + 1)/\text{pr}(Y = 0|\boldsymbol{X} + 1)}{\text{pr}(Y = 1|\boldsymbol{X})/\text{pr}(Y = 0|\boldsymbol{X})}\right\} \,,$$

representing the log-odds ratio for an individual to get disease if the risk factor is increased by one unit.

The logistic model indicates that the design is a cohort study that collects data forward to see how the disease develops. In the model risk factor $\boldsymbol{X}$ is regarded as a fixed quantity and Y is a random variable. In a case-control study, information is collected on the basis of disease status. However, the logistic model has an identity of inferential procedures about the log-odds ratio regardless of sampling approach whether it is carried out to a cohort or a case-control study.

The odds for getting disease for an individual with risk factor $\boldsymbol{X}$, relative to that for an individual with some reference risk factor $\boldsymbol{X}_0$ is

$$\frac{\text{pr}(Y = 1|\boldsymbol{X})/\text{pr}(Y = 0|\boldsymbol{X})}{\text{pr}(Y = 1|\boldsymbol{X}_0)/\text{pr}(Y = 0|\boldsymbol{X}_0)} = \exp\{(\boldsymbol{X} - \boldsymbol{X}_0)^{\mathrm{T}}\boldsymbol{\beta}\} \,, \tag{2.3}$$

by the Bayes rule, it is equivalent to

$$\frac{\text{pr}(\boldsymbol{X}|Y=1)/\text{pr}(\boldsymbol{X}_0|Y=1)}{\text{pr}(\boldsymbol{X}|Y=0)/\text{pr}(\boldsymbol{X}_0|Y=0)} = \exp\{(\boldsymbol{X}-\boldsymbol{X}_0)^{\mathrm{T}}\boldsymbol{\beta}\}\,.$$

Thus the odds ratio (2.3) can be estimated from case-control data (Prentice and Pyke, 1979), and risk factors can be modeled by an ordinary logistic regression model

$$\text{pr}(\boldsymbol{X}|Y=1) = \text{pr}(\boldsymbol{X}_0|Y=1)\exp\{\alpha^* + (\boldsymbol{X}-\boldsymbol{X}_0)^{\mathrm{T}}\boldsymbol{\beta}\}\,,$$

where $\alpha^* = \log\{\text{pr}(\boldsymbol{X}|Y=0)/\text{pr}(\boldsymbol{X}_0|Y=0)\}$.

In a case-control study with n_1 cases, n_0 controls and $n = n_0 + n_1$, suppose that z is the indicator whether a subject is sampled ($z=1$) or not ($z=0$), and

$$\pi_1 = \text{pr}(z=1|Y=1)$$

is the probability that a diseased subject is included in the study as a case and

$$\pi_0 = \text{pr}(z=1|Y=0)$$

is the probability that a disease-free subject is included in the study as a control. Under the assumption that sampling probability depend only on disease status and not on the risk factors, i.e.,

$$\text{pr}(z=1|Y=i,\boldsymbol{X}) = \text{pr}(z=1|Y=i)$$

for $i=0,1$, the conditional probability that a subject has disease given he has risk factor

$\boldsymbol{X}$ and he is sampled for the case-control study is

$$\begin{aligned} \text{pr}(Y=1|\boldsymbol{X}, z=1) &= \frac{\text{pr}(z=1|Y=1,\boldsymbol{X})\text{pr}(Y=1|\boldsymbol{X})}{\sum_{i=0}^{1}\text{pr}(z=1|Y=i,\boldsymbol{X})\text{pr}(Y=i|\boldsymbol{X})} \\ &= \frac{\pi_1 \exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})}{\pi_0+\pi_1\exp(\alpha+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})} \\ &= \frac{\exp(\alpha^*+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\alpha^*+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})}\,, \end{aligned} \tag{2.4}$$

where $\alpha^* = \alpha + \log(\pi_1/\pi_0)$.

From (2.4), for $i = 0$ and 1 we have

$$\begin{aligned} \text{pr}(\boldsymbol{X}|Y=i) &= \text{pr}(\boldsymbol{X}|Y=i, z=1) \\ &= \frac{\text{pr}(Y=i|\boldsymbol{X}, z=1)\text{pr}(\boldsymbol{X}|z=1)}{\text{pr}(Y=i|z=1)}\,, \end{aligned}$$

therefore the likelihood function of case-control study with risk factor $\boldsymbol{X}$ is

$$L(\boldsymbol{\beta}) = \prod_{j:cases}\text{pr}(\boldsymbol{X}_j|Y=1)\times\prod_{j:controls}\text{pr}(\boldsymbol{X}_j|Y=0) \;\propto L_1\times L_2\,,$$

where

$$L_1(\boldsymbol{\beta}) = \prod_{j:cases}\text{pr}(Y=1|\boldsymbol{X}_j, z=1)\times\prod_{j:controls}\text{pr}(Y=0|\boldsymbol{X}_j, z=1)$$

and

$$L_2 = \prod_{j=1}^{n}\text{pr}(\boldsymbol{X}_j|z=1)\,.$$

In most practical situations, the $\boldsymbol{X}$ variable does not contain any information of parameter $\boldsymbol{\beta}$, usually $\text{pr}(\boldsymbol{X}|z=1)$ is allowed to take arbitrary distribution, or may dependent a set of parameters that are independent of $\boldsymbol{\beta}$. Under this assumption, the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ obtained by maximizing L is the same as maximizing L_1 (Prentice and Pyke,

1979). Furthermore, $\widehat{\boldsymbol{\beta}}$ is consistent for log-odds ratio parameter $\boldsymbol{\beta}$ and also asymptotically normal distributed.

2.5 Logistic Regression in Matched Case-Control Studies

Suppose that data consist of n strata and there are M_i controls matched with a case for stratum $\boldsymbol{S}_i$, $i = 1, 2, \cdots, n$. Let Y_{ij} take on value one or zero according as the j^{th} subject in the i^{th} matched set is a case or control respectively, and $\boldsymbol{X}_{ij}$ be a vector of covariates. The prospective logistic regression for the chance of getting disease with risk factor $\boldsymbol{X}_{ij}$ is

$$\text{pr}(Y_{ij} = 1|\boldsymbol{S}_i, \boldsymbol{X}_{ij}) = H(\alpha_i + \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}),$$

where α_i is the stratum intercept term. Similarly to the general case-control study, the probability of a case is still included in the case-control study is modeled by

$$\text{pr}(Y_{ij} = 1|\boldsymbol{S}_i, \boldsymbol{X}_{ij}, z = 1) = H(\delta(\boldsymbol{S}_i) + \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}),$$

where $\delta(\boldsymbol{S}_i) = \alpha_i + \log(n_1/n_0)$.

The likelihood function for a matched case-control study satisfies

$$L(\boldsymbol{\beta}) \propto \prod_{i=1}^{n}\prod_{j=1}^{M_i} \text{pr}(Y_{ij}|\boldsymbol{S}_i, \boldsymbol{X}_{ij}, z = 1).$$

Note that the number of nuisance parameters $\delta(\boldsymbol{S}_i)$ is proportional to the number of strata which may cause inconsistent estimator of the log-odds ratio. To solve this problem, we consider the conditional likelihood function (Liddell et al., 1977; Breslow et al., 1978)

$$\begin{aligned} L_{\mathrm{C}}(\boldsymbol{\beta}) &= \prod_{i=1}^{n}\prod_{j=1}^{M_i+1} \text{pr}(Y_{ij}|\boldsymbol{S}_i, \boldsymbol{X}_{ij}, z = 1, \sum_{j=1}^{M_i+1} Y_{ij} = 1) \\ &= \prod_{i=1}^{n}\sum_{j=1}^{M_i+1} p_{ij}Y_{ij}, \end{aligned}$$

where $p_{ij} = \exp(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) / \sum_{j=1}^{M+i+1} \exp(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$.

The standard conditional maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ is obtained by maximizing the conditional likelihood function $L_{\mathrm{C}}(\boldsymbol{\beta})$.

CHAPTER III

BIAS REDUCTION IN CONDITIONAL LOGISTIC REGRESSION MODELS

3.1 Introduction

Conditional likelihood is widely used for the estimation of log-odds ratio from matched case-control studies. In the conditional likelihood analysis of a matched design, we remove the stratum specific nuisance parameters by conditioning on their sufficient statistics, and obtain consistent estimates of the log-odds ratio parameter by maximizing the conditional likelihood. The concern is that the maximum conditional likelihood (MCL) estimator obtained by maximizing the conditional likelihood is biased for small to moderate sample sizes. Since a matched case-control study with a small or moderate sample size is not uncommon, it is important to develop a method which can produce an estimate of the parameters of interest with little or less bias. For instance, the data example that motivates our research is from a low-birth-weight study discussed in Chapter I. It consists of only 29 strata, and each stratum has only one case and three controls.

In order to correct the bias of the log-odds ratio estimator based on matched studies different approaches have been proposed so far. There are two main types of approaches to handle the bias: a bias *corrective* method and a bias *preventive* method. Within the bias corrective methods, Jewell (1984) considered a computationally-intensive jackknife method for correcting the bias in the odds-ratio estimator for a categorical exposure variable and he compared the performance of his proposed method with some other bias correction approaches. He exclusively focused on a single exposure variable with finite many categories. Several other bias correction techniques for the odds-ratio estimator for discrete exposure variables were proposed by Bishop and Holland (1975). Cordeiro and McCullagh (1991) derived a general formula for the first-order term of bias that can be used in generalized

linear models. Greenland (2000) adopted this general bias correction approach to correct the bias in the conditional logistic regression setting. His method of bias correction in the log-odds ratio estimator is able to handle any types of exposures, and can easily handle multiple exposure variables simultaneously. In general we can describe the bias *corrective* method as follows. Let $\widehat{\boldsymbol{\beta}}$ be the MCL estimator of the parameter of interest $\boldsymbol{\beta}$, then the bias of the MCL estimator is

$$b(\boldsymbol{\beta}) = \frac{b_1(\boldsymbol{\beta})}{n} + \frac{b_2(\boldsymbol{\beta})}{n^2} + \cdots .$$

Thus the first-order bias corrected estimator for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{bc}} = \widehat{\boldsymbol{\beta}} - \frac{b_1(\widehat{\boldsymbol{\beta}})}{n}.$$

One of the drawbacks of the corrective approach is that if the MCL estimate $\widehat{\boldsymbol{\beta}}$ has an infinite component for a dataset, which is not uncommon for small sample sizes, then the first-order bias can be infinite in which cases the bias corrected estimator would be undefined. Note that the jackknife method of bias correction is even more likely to encounter the same problem. To avoid this difficulty Firth (1993) proposed a second type approach – a bias *preventive* method. This approach eliminates the first-order of bias $O(n^{-1})$ by solving a modified score equation in the general context of regular parametric families of distributions. Therefore the resultant estimator does not depend on the finiteness of the standard maximum likelihood estimator. Firth's idea has been used in the unconditional logistic regression (Heinze and Schemper, 2001b; Bull et al., 2002; Bull et al., 2007) and Cox's partial likelihood setting to handle problem of monotone likelihood (Heinze and Schemper, 2001a). We will use this bias preventive procedure in the context of conditional logistic regression models and derive the modified score function to estimate the model parameters. We refer to the method as a modified score function (MDS) approach.

This chapter is organized as follows. Section 3.2 contains the model and assumptions.

We present the proposed method in Section 3.3. We also derive a closed form expression for the bias-reduced estimator and its standard error for matched pair data with one dichotomous exposure variable. Section 3.4 contains a simulation study, where we compare the performance of the proposed method mainly with conditional maximum likelihood method and the jackknife approach of bias correction. An application in a low-birth-weight dataset is illustrated in Section 3.5. The discussion is given in Section 3.6.

In concluding this section we would like to highlight the novel features of this work. Firstly, to the best of our knowledge this is the first bias reduction approach in the MCL estimator in matched case-control studies using the preventive method. Secondly, compared to the jackknife or bootstrap based procedures the MDS method takes much less computational time and efforts. Thirdly, the MDS estimators are usually finite even when MCL estimates are infinite. Fourthly, we provide a versatile formula for standard error calculation for the MDS estimators.

3.2 Model and Assumptions

Suppose we have a 1:M_i (≥ 1) matched case-control data with n strata. Let Y_{ij} take on value one or zero according as the j^{th} subject in the i^{th} matched set is a case or control respectively. Let $\boldsymbol{X}_{ij} = (X_{ij1}, \cdots, X_{ijp})^{\mathrm{T}}$ be a $p \times 1$ vector of covariates. Also, let $\boldsymbol{S}_i$ be the covariates which are used for matching purposes in the i^{th} stratum. The disease risk in the i^{th} stratum can be modeled by

$$\text{pr}(Y_{ij} = 1 | \boldsymbol{S}_i, \boldsymbol{X}_{ij}) = H(\alpha_i(\boldsymbol{S}_i) + \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) \,, \tag{3.1}$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$, $j = 1, \cdots, M_i + 1$, and $i = 1, \cdots, n$. Note that α_i is the stratum specific parameter which is a function of $\boldsymbol{S}_i$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ is the vector of log-odds ratio parameters for the covariate $\boldsymbol{X}$.

In order to estimate $\boldsymbol{\beta}$ in Equation (3.1) one generally adopts the conditional logistic regression (Breslow and Day, 1980) where the estimates are obtained by maximizing

$$L_{\mathrm{C}}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \sum_{j=1}^{M_i+1} p_{ij} Y_{ij} \,, \tag{3.2}$$

where $p_{ij} = \exp(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) / \sum_{k=1}^{M_i+1} \exp(\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta})$, representing the conditional probability that the j^{th} subject is a case given that there is one case in the i^{th} stratum.

Notice that Y_{ij} takes on value one or zero, the likelihood function can be rewritten as

$$\begin{aligned} L_{\mathrm{C}}(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \sum_{j=1}^{M_i+1} p_{ij} Y_{ij} = \prod_{i=1}^{n} \prod_{j=1}^{M_i+1} p_{ij}^{Y_{ij}} \\ &= \prod_{i=1}^{n} \frac{\prod_{j=1}^{M_i+1} \exp(Y_{ij} \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})}{\sum_{k=1}^{M_i+1} \exp(\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta})} \,. \end{aligned}$$

Thus the log likelihood function

$$\begin{aligned} l_{\mathrm{C}}(\boldsymbol{\beta}) &= \log(L_{\mathrm{C}}(\boldsymbol{\beta})) \\ &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{M_i+1} Y_{ij} \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \log\left(\sum_{k=1}^{M_i+1} \exp(\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta}) \right) \right\} \,. \end{aligned}$$

The MCL estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ can obtained by solving the score equation

$$\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{0} \,,$$

where the score function

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial l_{\mathrm{C}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{j=1}^{M_i+1} (Y_{ij} - p_{ij}) \boldsymbol{X}_{ij} \,.$$

The variance of MCL estimator $\widehat{\boldsymbol{\beta}}$ can be estimated by the diagonal elements of $\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})$.

Since

$$\begin{aligned}\frac{\partial p_{ij}}{\partial \boldsymbol{\beta}} &= \frac{\exp(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{X}_{ij}\sum_{k=1}^{M_i+1}\exp(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) - \exp(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\sum_{k=1}^{M_i+1}\exp(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{X}_{ij}}{\left\{\sum_{k=1}^{M_i+1}\exp(\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta})\right\}^2} \\ &= p_{ij}\left\{\boldsymbol{X}_{ij} - \sum_{k=1}^{M_i+1} p_{ik}\boldsymbol{X}_{ik}\right\},\end{aligned}$$

the $\boldsymbol{I}(\boldsymbol{\beta})$ which is the Fisher's information matrix is given by

$$\begin{aligned}\boldsymbol{I}(\boldsymbol{\beta}) &= -E\left(\frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right) \\ &= \sum_{i=1}^{n}\left\{\sum_{j=1}^{M_i+1} p_{ij}\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^{\mathrm{T}} - \left(\sum_{j=1}^{M_i+1}\boldsymbol{X}_{ij}p_{ij}\right)\left(\sum_{j=1}^{M_i+1}\boldsymbol{X}_{ij}p_{ij}\right)^{\mathrm{T}}\right\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{M_i+1}\left(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{i\cdot}\right)\left(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{i\cdot}\right)^{\mathrm{T}} p_{ij},\end{aligned}$$

where $\overline{\boldsymbol{X}}_{i\cdot} = \sum_{j=1}^{M_i+1} p_{ij}\boldsymbol{X}_{ij}$.

We now mention two special data configurations: complete separation and quasicomplete separation which are defined as follows. Following Albert and Anderson (1984) we say a 1:M_i matched case-control data is completely separated if there exists a vector $\boldsymbol{\gamma} \in R^p$, and $\boldsymbol{\gamma} \neq \mathbf{0}$, such that $\boldsymbol{\gamma}^T(\boldsymbol{X}_{i1} - \boldsymbol{X}_{ij}) > 0$ for all $j = 2, \cdots, M_i + 1$ and $i = 1, \cdots, n$, assuming that $Y_{i1} = 1$ and $Y_{ij} = 0$. If a 1:M_i matched case-control data is not completely separated but there exists a $\boldsymbol{\gamma} \neq \mathbf{0}$ such that $\boldsymbol{\gamma}^T(\boldsymbol{X}_{i1} - \boldsymbol{X}_{ij}) \geq 0$ (i.e., the equality holds necessarily for at least one (i, j)), then we call it quasicompletely separated. We have the following Lemma:

LEMMA: If matched case-control data are completely or quasicompletely separated, the MCL estimate is infinite.

Proof. In this proof we follow the idea of Albert and Anderson (1984). For notational convenience we will denote $\boldsymbol{X}_{ij} - \boldsymbol{X}_{i1}$ by $\boldsymbol{Z}_{ij}$. Let $\mathcal{B}$ be the set of vectors such that for any $\boldsymbol{\gamma} \in \mathcal{B}$, $\boldsymbol{\gamma}^T\boldsymbol{Z}_{ij} \leq 0$, for $j = 2, \cdots, M_i$, and there exists at least one (i, j) for

which the equality holds. Observe that $\mathcal{B}$ is a non-empty convex set for a quasi-separated dataset. For any $\boldsymbol{\beta} \in \mathcal{R}^p$, we can find a projection of it on $\mathcal{B}$, $\boldsymbol{\gamma}$, such that $\boldsymbol{\beta} = \boldsymbol{\gamma} + \boldsymbol{\alpha}$, $\boldsymbol{\alpha} \in \mathcal{B}^c$. Let $\mathcal{D}_i(\boldsymbol{\gamma}) \equiv \{j : \boldsymbol{\gamma}^T \boldsymbol{Z}_{ij} = 0\}$ and $\mathcal{D}_i^c(\boldsymbol{\gamma}) \equiv \{j : \boldsymbol{\gamma}^T \boldsymbol{Z}_{ij} < 0\}$, and define $Q(\boldsymbol{\gamma}) \equiv \{i : |\mathcal{D}_i(\boldsymbol{\gamma})| > 0\}$, where $|\mathcal{D}_i|$ is the cardinality of $\mathcal{D}_i$. Furthermore, define $\boldsymbol{\beta}(k) = k\boldsymbol{\gamma} + \boldsymbol{\alpha}$, for $k > 0$. Now,

$$
\begin{aligned}
L_C\{\boldsymbol{\beta}(k)\} = & \prod_{i \in Q(\boldsymbol{\gamma})} \frac{1}{1 + \sum_{j \in \mathcal{D}_i(\boldsymbol{\gamma})} \exp(k\boldsymbol{\gamma}^T \boldsymbol{Z}_{ij} + \boldsymbol{\alpha}^T \boldsymbol{Z}_{ij}) + \sum_{j \in \mathcal{D}_i^c(\boldsymbol{\gamma})} \exp(k\boldsymbol{\gamma}^T \boldsymbol{Z}_{ij} + \boldsymbol{\alpha}^T \boldsymbol{Z}_{ij})} \\
& \times \prod_{i \in Q^c(\boldsymbol{\gamma})} \frac{1}{1 + \sum_j \exp(k\boldsymbol{\gamma}^T \boldsymbol{Z}_{ij} + \boldsymbol{\alpha}^T \boldsymbol{Z}_{ij})} \\
= & \prod_{i \in Q(\boldsymbol{\gamma})} \frac{1}{1 + \sum_{j \in \mathcal{D}_i(\boldsymbol{\gamma})} \exp(\boldsymbol{\alpha}^T \boldsymbol{Z}_{ij}) + \sum_{j \in \mathcal{D}_i^c(\boldsymbol{\gamma})} \exp(k\boldsymbol{\gamma}^T \boldsymbol{Z}_{ij} + \boldsymbol{\alpha}^T \boldsymbol{Z}_{ij})} \\
& \times \prod_{i \in Q^c(\boldsymbol{\gamma})} \frac{1}{1 + \sum_j \exp(k\boldsymbol{\gamma}^T \boldsymbol{Z}_{ij} + \boldsymbol{\alpha}^T \boldsymbol{Z}_{ij})}.
\end{aligned}
$$

Note that $L_C\{\boldsymbol{\beta}(k)\}$ is a monotonically increasing function in k, and the supremum attains when $k \to \infty$, that is,

$$
\sup_k L_C\{\boldsymbol{\beta}(k)\} = \prod_{i \in Q(\boldsymbol{\gamma})} \{1 + \sum_{j \in \mathcal{D}_i(\boldsymbol{\gamma})} \exp(\boldsymbol{\alpha}^T \boldsymbol{Z}_{ij})\}^{-1} \text{ as } k \to \infty \,.
$$

Thus, $L_C(\boldsymbol{\beta}) = L_C\{\boldsymbol{\beta}(1)\} < \sup_k L_C\{\boldsymbol{\beta}(k)\}$ for every $\boldsymbol{\beta} \in \mathcal{R}^p$. Therefore, the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ must be infinity, since if $\widehat{\boldsymbol{\beta}}$ were finite, then the argument above indicates that there is a large enough K, such that $L_C\{\widehat{\boldsymbol{\beta}}(k)\} > L_C(\widehat{\boldsymbol{\beta}})$ for $k \geq K$, leading to a contradiction.

Similarly, the conclusion holds true when the data are completely separated for which $Q(\boldsymbol{\gamma})$ is an empty set, and then the $\sup_k L_C\{\boldsymbol{\beta}(k)\} = 1$.

3.3 Method of Bias Reduction

Firth (1993) gave a geometric interpretation of the modified score function. The basic idea is that the bias in $\widehat{\boldsymbol{\beta}}$ can be reduced by introducing a small bias in $\boldsymbol{U}(\boldsymbol{\beta})$. By Taylor's series expansion we can write

$$\begin{aligned}\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) &= \boldsymbol{U}(\boldsymbol{\beta}) + \frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \cdots \\ &= \left\{\boldsymbol{U}(\boldsymbol{\beta}) + \frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\frac{b_1(\boldsymbol{\beta})}{n}\right\} + \frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\left\{\widehat{\boldsymbol{\beta}} - \frac{b_1(\boldsymbol{\beta})}{n} - \boldsymbol{\beta}\right\} + \cdots \\ &= \mathbf{0}\,,\end{aligned}$$

where $b_1(\boldsymbol{\beta}) = (b_{1(1)}(\boldsymbol{\beta}), \cdots, b_{1(p)}(\boldsymbol{\beta}))^{\mathrm{T}}$. This expansion indicates that the estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{R}}$ obtained by solving the modified score

$$\boldsymbol{U}^{\mathrm{mod}}(\boldsymbol{\beta}) = \boldsymbol{U}(\boldsymbol{\beta}) + \frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\frac{b_1(\boldsymbol{\beta})}{n} = \mathbf{0}$$

yields a first-order bias preventive estimator of $\boldsymbol{\beta}$. Observe that the bias reduction is implicit through the modified score. Therefore when obtaining the bias reduced estimator we do not need to subtract an estimated bias from the original MCL estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. This is the key idea from Firth (1993) used in our bias reduction approach in matched case-control studies. Using the notation $\boldsymbol{I}(\boldsymbol{\beta}) = -\partial \boldsymbol{U}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^{\mathrm{T}}$, we can rewrite the modified score as

$$\boldsymbol{U}^{\mathrm{mod}}(\boldsymbol{\beta}) = \boldsymbol{U}(\boldsymbol{\beta}) - \boldsymbol{I}(\boldsymbol{\beta})\frac{b_1(\boldsymbol{\beta})}{n}\ .$$

To describe our methodology more effectively, we define

$$\begin{aligned}k_{rs} &= \frac{1}{n}E\left\{\frac{\partial^2 l_{\mathrm{C}}(\boldsymbol{\beta})}{\partial\beta_r\partial\beta_s}\right\}\,, \\ k_{rst} &= \frac{1}{n}E\left\{\frac{\partial^3 l_{\mathrm{C}}(\boldsymbol{\beta})}{\partial\beta_r\partial\beta_s\partial\beta_t}\right\}\,, \\ k_{rs,t} &= \frac{1}{n}E\left\{\left[\frac{\partial^2 l_{\mathrm{C}}(\boldsymbol{\beta})}{\partial\beta_r\partial\beta_s}\right]\cdot\left[\frac{\partial l_{\mathrm{C}}(\boldsymbol{\beta})}{\partial\beta_t}\right]\right\}\ .\end{aligned}$$

Let $\boldsymbol{D} = \{k_{rs}\}$ and its inverse $\boldsymbol{D}^{-1} = \{k^{rs}\}$. Then following Cordeiro and McCullagh (1991) and using the fact that $k_{rs,t} = 0$, the first-order bias term for β_r is

$$\frac{b_{1(r)}(\boldsymbol{\beta})}{n} = \frac{1}{2n}\sum_{a=1}^{p}\sum_{t=1}^{p}\sum_{u=1}^{p} k^{ra}k^{tu}k_{atu}$$

for $r = 1, \cdots, p$.

Define the conventional notations such as

$$k_{r,s} = \frac{1}{n}E\left\{\left[\frac{\partial l_{\mathrm{C}}(\boldsymbol{\beta})}{\partial \beta_r}\right]\cdot\left[\frac{\partial l_{\mathrm{C}}(\boldsymbol{\beta})}{\partial \beta_s}\right]\right\}$$

and

$$k_{r,s,t} = \frac{1}{n}E\left\{\left[\frac{\partial l_{\mathrm{C}}(\boldsymbol{\beta})}{\partial \beta_r}\right]\cdot\left[\frac{\partial l_{\mathrm{C}}(\boldsymbol{\beta})}{\partial \beta_s}\right]\cdot\left[\frac{\partial l_{\mathrm{C}}(\boldsymbol{\beta})}{\partial \beta_t}\right]\right\},$$

the r^{th} component of $-\boldsymbol{I}(\boldsymbol{\beta})b_1(\boldsymbol{\beta})/n$ is then

$$c_r = \frac{1}{2}\sum_{s=1}^{p}\sum_{a=1}^{p}\sum_{b=1}^{p}\sum_{c=1}^{p} k_{rs}k^{sa}k^{bc}k_{abc} = \frac{1}{2}\sum_{b=1}^{p}\sum_{c=1}^{p} k^{bc}k_{rbc}$$

as $\sum_{s=1}^{p} k_{rs}k^{sa} = \boldsymbol{I}(r = a)$ for $r = 1, \cdots, p$.

Following the identities $k_{rst} + k_{r,st} + k_{s,rt} + k_{t,rs} + k_{r,s,t} = 0$ and $k_{r,s} + k_{rs} = 0$ we can write the r^{th} element of the correction terms as

$$c_r = \frac{1}{2}\frac{\partial}{\partial \beta_r}\Big\{\log|\boldsymbol{I}(\boldsymbol{\beta})|\Big\}.$$

Letting $\boldsymbol{U}^{\mathrm{mod}}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\boldsymbol{U}_i^{\mathrm{mod}}(\boldsymbol{\beta}) = (\boldsymbol{U}_{(1)}^{\mathrm{mod}}(\boldsymbol{\beta}), \cdots, \boldsymbol{U}_{(p)}^{\mathrm{mod}}(\boldsymbol{\beta}))^{\mathrm{T}}$ and $\boldsymbol{U}(\boldsymbol{\beta}) =$

$(\boldsymbol{U}_{(1)}(\boldsymbol{\beta}), \cdots, \boldsymbol{U}_{(p)}(\boldsymbol{\beta}))^{\mathrm{T}}$, the r^{th} component of the modified score function is

$$\begin{aligned} U_{(r)}^{\mathrm{mod}}(\boldsymbol{\beta}) &= \boldsymbol{U}_{(r)}(\boldsymbol{\beta}) + \frac{1}{2}\frac{\partial}{\partial\beta_r}\left\{\log|\boldsymbol{I}(\boldsymbol{\beta})|\right\} \\ &= \boldsymbol{U}_{(r)}(\boldsymbol{\beta}) + \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{I}^{-1}(\boldsymbol{\beta})\frac{\partial \boldsymbol{I}(\boldsymbol{\beta})}{\partial\beta_r}\right\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{M_i+1}(Y_{ij}-p_{ij})X_{ijr} + \frac{1}{2}\mathrm{tr}\Bigg[\left\{\sum_{i=1}^{n}\sum_{j=1}^{M_i+1}(\boldsymbol{X}_{ij}-\overline{\boldsymbol{X}}_{i\cdot})(\boldsymbol{X}_{ij}-\overline{\boldsymbol{X}}_{i\cdot})^{\mathrm{T}}p_{ij}\right\}^{-1} \\ &\quad \times \sum_{i=1}^{n}\sum_{j=1}^{M_i+1}(\boldsymbol{X}_{ij}-2\overline{\boldsymbol{X}}_{i\cdot})\boldsymbol{X}_{ij}^{\mathrm{T}}(X_{ijr}-\overline{X}_{i\cdot r})p_{ij}\Bigg], \end{aligned}$$

where $\overline{X}_{i\cdot r} = \sum_{j=1}^{M_i+1} X_{ijr}p_{ij}$. Following the argument in Firth (1993), it is seen that the adjustment term after the standard conditional score function above effectively eliminates the first-order bias of the conditional maximum likelihood estimator.

Note that we can write the r^{th} component of the modified score function $U_{(r)}^{\mathrm{mod}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} U_{(r)i}^{\mathrm{mod}}(\boldsymbol{\beta})$, where $U_{(r)i}^{\mathrm{mod}}(\boldsymbol{\beta}) = \sum_{j=1}^{M_i+1}(Y_{ij}-p_{ij})\boldsymbol{X}_{ij} - \boldsymbol{I}_{(r)}(\boldsymbol{\beta})b_1(\boldsymbol{\beta})/n^2$, and $\boldsymbol{I}_{(r)}(\boldsymbol{\beta})$ is the r^{th} row of $\boldsymbol{I}(\boldsymbol{\beta})$.

Let $\boldsymbol{U}^{\mathrm{mod}}(\boldsymbol{\beta}) = (U_{(1)}^{\mathrm{mod}}(\boldsymbol{\beta}), \cdots, U_{(p)}^{\mathrm{mod}}(\boldsymbol{\beta}))^{\mathrm{T}}$. Since we proposed a new equation to calculate the estimator, naturally we need to estimate its asymptotic variance. Now expanding $\boldsymbol{U}^{\mathrm{mod}}(\widehat{\boldsymbol{\beta}}_{\mathrm{R}})$ at the true $\boldsymbol{\beta}$,

$$\begin{aligned} \mathbf{0} &= n^{-1/2}\boldsymbol{U}^{\mathrm{mod}}(\widehat{\boldsymbol{\beta}}_{\mathrm{R}}) \\ &= n^{-1/2}\sum_{i=1}^{n}\boldsymbol{U}_i^{\mathrm{mod}}(\boldsymbol{\beta}) + n^{-1/2}\frac{\partial \boldsymbol{U}^{mod}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}}_{\mathrm{R}}-\boldsymbol{\beta}) + \boldsymbol{O}_p(n^{-1/2}), \end{aligned}$$

we have

$$\widehat{\boldsymbol{\beta}}_{\mathrm{R}} = \boldsymbol{\beta} + \left(\frac{\partial \boldsymbol{U}^{mod}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right)^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i^{\mathrm{mod}}(\boldsymbol{\beta}) + \boldsymbol{O}_p(n^{-1}).$$

Thus the sandwich typed estimator of the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{R}}$ is

given by

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_{\mathrm{R}}) = \left[\left(\frac{\partial \boldsymbol{U}^{\mathrm{mod}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{U}_i^{\mathrm{mod}}(\boldsymbol{\beta})\boldsymbol{U}_i^{\mathrm{mod}\,\mathrm{T}}(\boldsymbol{\beta})\right)\left(\frac{\partial \boldsymbol{U}^{\mathrm{mod}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right)^{-\mathrm{T}}\right]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\mathrm{R}}}, \quad (3.3)$$

where

$$\frac{\partial U_{(r)}^{\mathrm{mod}}(\boldsymbol{\beta})}{\partial \beta_k} = \frac{\partial U_{(r)}(\boldsymbol{\beta})}{\partial \beta_k} + \frac{1}{2}\mathrm{tr}\left\{-I^{-1}(\boldsymbol{\beta})\frac{\partial I(\boldsymbol{\beta})}{\partial \beta_k}I^{-1}(\boldsymbol{\beta})\frac{\partial I(\boldsymbol{\beta})}{\partial \beta_r} + I^{-1}(\boldsymbol{\beta})\frac{\partial^2 I(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_r}\right\}.$$

Here

$$\frac{\partial I(\boldsymbol{\beta})}{\partial \beta_r} = \sum_{i=1}^{n}\sum_{j=1}^{M_i+1}(\boldsymbol{X}_{ij} - 2\bar{\boldsymbol{X}}_{i\cdot})\boldsymbol{X}_{ij}^{\mathrm{T}}(X_{ijr} - \bar{X}_{i\cdot r})p_{ij}$$

and

$$\begin{aligned}\frac{\partial^2 I(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_r} = \sum_{i=1}^{n}\sum_{j=1}^{M_i+1}\Bigg[&\left(-2\frac{\partial \bar{\boldsymbol{X}}_{i\cdot}}{\partial \beta_k}\boldsymbol{X}_{ij}^{\mathrm{T}}\right)(X_{ijr} - \bar{X}_{i\cdot r})p_{ij} + \\ &(\boldsymbol{X}_{ij} - 2\bar{\boldsymbol{X}}_{i\cdot})\boldsymbol{X}_{ij}^{\mathrm{T}}\left\{(X_{ijr} - \bar{X}_{i\cdot r})\frac{\partial p_{ij}}{\beta_k} - \frac{\partial \bar{X}_{i\cdot r}}{\beta_k}p_{ij}\right\}\Bigg]\end{aligned}$$

with $\partial p_{ij}/\partial \beta_k = p_{ij}(X_{ijk} - \bar{X}_{i\cdot k})$, $\partial \bar{X}_{i\cdot r}/\partial \beta_k = \sum_{j=1}^{M_i+1} X_{ijr}(\partial p_{ij}/\partial \beta_k)$, and $\partial \bar{\boldsymbol{X}}_{i\cdot}/\partial \beta_k = \sum_{j=1}^{M_i+1} \boldsymbol{X}_{ij}(\partial p_{ij}/\partial \beta_k)$.

Although according to the development in Firth (1993) the modified score estimator has the same asymptotic variance-covariance matrix as the maximum conditional likelihood estimator does, in the simulation study we found that for small to moderate sample sizes the sandwich type estimator (3.3) yields more accurate estimate of the true standard error than that obtained by inverting the Fisher's information matrix. One explanation for more accuracy of formula (3.3) is that it takes into account the correction term of the modified score function which is of the order $O(1)$, and the effect of this correction term on the variance may not be negligible for small n.

Here we briefly mention the connection between this bias reduction approach and the

Bayesian approach. Often for small sample sizes, Bayesian method with a prior distribution of good faith can circumvent the problem of bias, and usually the prior belief depends on the historical studies. However, the concern is how one can proceed with the Bayesian method in the absence of precise prior knowledge about the parameter. Firth (1993) showed that the estimator obtained by the preventive method is actually the posterior mode of the parameter of interest with the Jeffrey's prior on the parameter, which shows that not only precise prior belief is useful, but also sometimes the objective prior can help us to remove the bias of the maximum likelihood estimator.

3.3.1 Case of a Single Covariate

For a single covariate the first-order bias-reduced estimator of the log-odds-ratio parameter is obtained by solving

$$\sum_{i=1}^{n}\sum_{j=1}^{M_i+1}(Y_{ij}-p_{ij})X_{ij}+\frac{\sum_{i=1}^{n}\sum_{j=1}^{M_i+1}X_{ij}(X_{ij}-\overline{X}_{i\cdot})(X_{ij}-2\overline{X}_{i\cdot})p_{ij}}{2\sum_{i=1}^{n}\sum_{j=1}^{M_i+1}(X_{ij}-\overline{X}_{i\cdot})^2p_{ij}}=0\,. \qquad (3.4)$$

The first term on the left hand side above is the score function derived from the conditional likelihood (5.3) and the second term is the correction term. To show the advantage of the MDS method compared to the MCL approach, now consider a $1{:}M_i$ matched study with $j=1$ representing the cases and otherwise controls. Assume that the data are completely separated, e.g., $X_{i1}>0$ and $X_{ij}<0$ for $j=2,\cdots,M_i+1$. The score equation derived from the conditional likelihood then is

$$\sum_{i=1}^{n}(1-p_{i1})X_{i1}=\sum_{i=1}^{n}\sum_{j=2}^{M_i+1}p_{ij}X_{ij}\,.$$

It is clearly seen that the left hand side of the equation is positive whereas the right hand side is negative. Thus there is no finite solution for β. In contrast when we use Equation (3.4) we do not experience such a problem in our numerical computations.

We now provide the closed form expression for the modified score function for a 1:M (≥ 1) matched case-control study with a dichotomous exposure variable. Let $n_{k,m,M}$ denote the number of matched sets containing M controls m of which are exposed and the case is ($k = 1$) or is not ($k = 0$) exposed. In addition, let $T_{m,M} = n_{1,m-1,M} + n_{0,m,M}$ be the number of such sets having a total of m exposed. Then the MCL estimator of β can be obtained as the solution of

$$\sum_{m=1}^{M} n_{1,m-1,M} = \sum_{m=1}^{M} T_{m,M} g_M(m,\beta)\,,$$

where $g_M(m,\beta) = m\exp(\beta)/\{(M+1-m)+m\exp(\beta)\}$ (Breslow and Day, 1980, p.177). Using the MDS method the bias-reduced estimator of β can be obtained as the solution to

$$\sum_{m=1}^{M} n_{1,m-1,M} - \sum_{m=1}^{M} T_{m,M} g_M(m,\beta) + \frac{\sum_{m=1}^{M} T_{m,M}\{(m+1-M)/m\}\{1-2g_M(m,\beta)\}g_M^2(m,\beta)}{2\sum_{m=1}^{M} T_{m,M}\{(m+1-M)/m\}g_M^2(m,\beta)} = 0\,.$$

One can easily compare the above modified conditional score with the conditional score function given in Equation (5.26) of Breslow and Day (1980, p. 177).

3.3.2 Matched Pair Design with a Dichotomous Exposure

Now we consider the matched pair design, i.e., $M_i = 1$ for $i = 1, 2, \cdots, n$, and with the exposure X taking on zero or one. Let u be the number of discordant matched pairs where the case is exposed and control is unexposed, i.e., $Y = 1, X = 1; Y = 0, X = 0$, and v be the number of discordant matched pairs where the case is unexposed and the control is exposed, i.e., $Y = 1, X = 0; Y = 0, X = 1$.

The MCL estimator of the log odds-ratio is

$$\widehat{\beta} = \log\left(\frac{u}{v}\right)$$

with variance

$$\mathrm{var}(\widehat{\beta}) = \frac{1}{(u+v)H(\beta)\{1-H(\beta)\}}\,.$$

A consistent estimator of $\mathrm{var}(\widehat{\beta})$ is

$$\widehat{\mathrm{var}}(\widehat{\beta}) = \frac{u+v}{uv}.$$

In this case Greenland's bias corrected estimator of β is given by

$$\widehat{\beta}_{\mathrm{G}} = \log\left(\frac{u}{v}\right) - \frac{(u-v)^2}{2uv(u+v)}.$$

No explicit formula for its asymptotic variance is given even though it can be approximated by the delta method as indicated in Greenland (2000).

In the MDS method we estimate β by solving

$$U^{\mathrm{mod}}(\beta) = U(\beta) + \frac{1}{2}\, I^{-1}(\beta)\frac{\partial I(\beta)}{\partial \beta} = 0\ ,$$

where

$$\begin{aligned} U(\beta) &= u-(u+v)H(\beta)\,, \\ I(\beta) &= (u+v)H(\beta)\{1-H(\beta)\}\,, \\ \frac{\partial I(\beta)}{\partial \beta} &= (u+v)H(\beta)\{1-H(\beta)\}\{1-2H(\beta)\}\,. \end{aligned}$$

The bias-reduced estimator becomes

$$\widehat{\beta}_{\mathrm{R}} = \log\left(\frac{2u+1}{2v+1}\right).$$

Note that for a small sample the MCL estimate could be infinite. Clearly, it is more likely for the jackknife estimator to have this difficulty. However, as we see $\widehat{\beta}_{\mathrm{R}}$ will not be infinite even for $v = 0$.

Furthermore,

$$\frac{\partial U^{\mathrm{mod}}(\beta)}{\partial \beta} = -(1+u+v)H(\beta)\{1-H(\beta)\}$$

and

$$\begin{aligned}\sum_{i=1}^{n}\{U^{\mathrm{mod}}_{(1)i}(\beta)\}^2 &= \frac{1}{4n}\{1-2H(\beta)\}^2 + u\{1-H(\beta)\}^2 + vH^2(\beta) \\ &\quad + \frac{1}{n}(u-v)H(\beta)\{1-H(\beta)\}\,.\end{aligned}$$

Then a consistent estimator of var$(\widehat{\beta}_{\mathrm{R}})$ is

$$\widehat{\mathrm{var}}(\widehat{\beta}_{\mathrm{R}}) = 4\left\{\frac{u}{(1+2u)^2} + \frac{v}{(1+2v)^2} - \frac{(u-v)^2}{n(1+2u)^2(1+2v)^2}\right\}.$$

Although, mathematically $\widehat{\mathrm{var}}(\widehat{\beta}_{\mathrm{R}}) < \widehat{\mathrm{var}}(\widehat{\beta}) = (u+v)/(uv)$, we cannot claim the superiority of one over the other. In addition, if one carries out a binomial experiment with $u+v$ trials, and u success are observed, then $\widehat{\beta}_{\mathrm{R}}$ becomes identical to the modified empirical estimator of the logit of the success probability of the binomial experiment given in Cox and Snell (1968, Section 2.1.9). This fact was also recognized by Firth (1993) in the context of unconditional logistic regression, and for conditional logistic regression for 1:1 matched case-control study as the later is equivalent to the unconditional logistic regression with appropriately defined covariates. Notice that the variance estimator $\widehat{\mathrm{var}}(\widehat{\beta}_{\mathrm{R}})$ is not the same as the variance estimator given in Cox and Snell (1968), however, if $u/(u+v) \to \rho$ as $n \to \infty$, then

$$\begin{aligned}\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}_u}) &= \frac{u(1+2v)^2 + v(1+2u)^2}{(1+2u)^2(1-2v)^2)} - \frac{(u-v)^2}{n(1+2u)^2(1+2v)^2} \\ &= \frac{1}{(u+v)\rho(1-\rho)} + O(n^{-1})\,.\end{aligned}$$

Thus $\widehat{\mathrm{var}}(\widehat{\beta}_{\mathrm{u}})/\{(u+v)\rho(1-\rho)\}^{-1} \to 1$, i.e., both variance estimators are asymptotically first-order equivalent as $n \to \infty$.

3.4 A Simulation Study

In order to judge the performance of the methods we conducted the following simulation study. We first generated a cohort data of with three variables S, X, and Y according to the simulation scenarios 1 and 2, and with four variables S, X, Z and Y according to simulation scenario 3.

1. 1:1 matched case-control data with $S \sim \text{Normal}(0.53, 0.24^2)$, $X \sim \text{Bernoulli}(p_x)$, $p_x = H(-2.0 + S)$, and marginally $p_x \approx 0.2$, and $Y \sim \text{Bernoulli}(p_y)$, where $p_y = H(\beta_0 + 1.1S + \beta X)$; with $\beta_0 = -2.9$ and $\beta = 0.5, 1.0$.

2. 1:2 matched case-control data with $S \sim \text{Normal}(0.53, 0.24^2)$, $X \sim \text{Gamma}(S^2 + 0.5, 2.5)$, and $Y \sim \text{Bernoulli}(p_y)$, where $p_y = H(\beta_0 + 1.1S + \beta X)$; with $\beta_0 = -3.0$ for $\beta = 0.5, 1.0$, $\beta_0 = -3.4$ for $\beta = 1.5$, and $\beta_0 = -3.7$ for $\beta = 2.0$.

3. 1:1 matched case-control data with $S \sim \text{Normal}(.53, 0.24^2)$, $Z \sim \text{Bernoulli}(p_z)$, $p_z = H(-1.7 + 1.3S)$, $X \sim \text{Bernoulli}(p_x)$, $p_x = H(-2.4 + 1.4S + 1.0Z)$, and $Y \sim \text{Bernoulli}(p_y)$, where $p_y = H(\beta_0 + 1.1S + \beta_1 Z + \beta_2 X)$; with $\beta_0 = -3.0$ for $(\beta_1, \beta_2) = (0.5, 0.5)$ and $\beta_0 = -3.3$ for $(\beta_1, \beta_2) = (1.0, 1.0)$.

The distribution and log-odds ratio parameter for S in p_y were chosen by mimicking a real data that we analyzed. We chose a value of β_0 so that the overall marginal disease prevalence is around 10%. The association between X and S was small to moderate, such as corr$(X, S) = 0.1, 0.28$, and 0.15 for the three scenarios, respectively.

From the cohort data we created $1{:}M$ matched case-control data with n strata using S as the matching variable. For all three scenarios we considered different values of n. Under each of the scenarios we generated $N = 2000$ datasets, and for each dataset three different estimates were obtained, the MCL estimate, the jackknife (JNF) estimate, and the MDS estimate.

In the jackknife method we treated each stratum as an individual unit, each time one stratum is leaving out and the rest of $n-1$ strata are used to calculate the new estimator. The process is repeated n times. Denote the new estimator after deleting the i^{th} stratum as $\widehat{\beta}_{(-i)}$, $i = 1, 2, \cdots, n$, the bias of jackknife estimator is given by $(n-1)\{\widehat{\beta}_{(\cdot)} - \widehat{\beta}\}$, where $\widehat{\beta}_{(\cdot)} = 1/n \sum_{i=1}^{n} \widehat{\beta}_{(-i)}$. Thus the jackknife bias corrective estimator is $\widehat{\beta}_J = n\widehat{\beta} - (n-1)\widehat{\beta}_{(\cdot)}$, and its variance is estimated by $\widehat{\text{var}}(\widehat{\beta}_J) = (n-1)/n \sum_{i=1}^{n} \{\widehat{\beta}_{(-i)} - \widehat{\beta}_{(\cdot)}\}^2$.

For the purpose of comparisons, we present the bias (with its empirical standard error), the estimated standard error (ESD), the 95% coverage probability (CP) based on a Wald-type confidence interval, and the "true" (empirical) standard error (TSD). In order to reduce the effect of some extreme observations we use the median absolute deviation (MAD): $\text{median}_{1 \le k \le N} |\widetilde{\beta}_k - \text{median}(\widetilde{\boldsymbol{\beta}})|/0.6745$ for TSD, where $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \widetilde{\beta}_2, \cdots, \widetilde{\beta}_N)^T$ and $\widetilde{\beta}_k$ represents the estimate for the k^{th} simulated dataset (Huber, 1981, p.144). In the simulation we deleted the datasets when the absolute value of the MCL or JNF estimates > 10 (in such cases the corresponding standard errors were always > 1000).

The summary quantities are calculated for all three approaches conditional on the datasets where both the MCL and JNF estimates are finite. In this case it is not surprising that MDS estimates often show bias because a particular data configuration which may occur with non-zero probability is discarded. However, the performance of a method should be judged based on all datasets instead of on its subset. From simulation results we see that in all scenarios we considered the MDS method does not produce infinite value of the estimators. Therefore we should evaluate the MDS method based on all 2000 replications. The results are also provided in the tables.

In the tables, the summary quantities for a method with an * refer to the approximated values obtained from the datasets where both MCL and JNF estimate are finite. In this context, the results in Tables 4 – 6 corresponding to scenarios 1 – 3, respectively, can be summarized as follows:

Table 4. Results of the simulation study for one binary covariate and $M = 1$. MCL, JNF and MDS stand for the maximum conditional likelihood, jackknife and the modified score estimators, respectively. TSD, ESD and CP represent the "true" standard error, estimated standard error, and nominal 95% coverage probability based on a Wald-type confidence interval. †: approximation by the MAD method; *: estimate was calculated based on the datasets where both MCL and JNF exist out of 2000 replications ; When $\beta = 0.5$ the number of divergent datasets out of 2000 replication are 165, 15, for $n = 30$ and 50 respectively, and when $\beta = 1$ the number of divergent datasets are 272, 40, 1 for $n = 30$, 50 and 100 respectively.

	$n = 30$				$n = 50$				$n = 100$			
Method	Bias	TSD†	ESD	CP	Bias	TSD†	ESD	CP	Bias	TSD†	ESD	CP
						$\beta = 0.5$						
MCL*	-0.041	0.697	0.666	0.988	0.017	0.510	0.516	0.965	0.004	0.347	0.356	0.964
JNF*	-0.118	0.597	0.770	0.996	-0.030	0.475	0.561	0.986	-0.014	0.341	0.367	0.974
MDS*	-0.087	0.645	0.589	0.973	-0.017	0.479	0.478	0.957	-0.012	0.341	0.343	0.959
MDS	-0.001	0.670	0.599	0.945	-0.004	0.499	0.479	0.952	-0.012	0.341	0.343	0.959
						$\beta = 1.0$						
MCL*	-0.110	0.529	0.658	0.975	0.018	0.505	0.524	0.977	0.024	0.349	0.364	0.955
JNF*	-0.248	0.418	0.772	0.985	-0.072	0.426	0.581	0.979	-0.012	0.330	0.380	0.975
MDS*	-0.194	0.499	0.578	0.958	-0.046	0.452	0.479	0.963	-0.007	0.336	0.349	0.951
MDS	-0.012	0.574	0.584	0.930	-0.014	0.443	0.482	0.951	-0.006	0.337	0.349	0.951

Table 5. Results of the simulation study for one continuous covariate and $M = 2$. MCL, JNF and MDS stand for the maximum conditional likelihood, jackknife and the modified score estimators, respectively. TSD, ESD and CP represent the "true" standard error, estimated standard error and nominal 95% confidence interval coverage probability. †: approximation by the MAD method; *: estimate was calculated based on the datasets where both MCL and JNF exist out of 2000 replications; When $\beta = 2$ the number of divergent dataset out of 2000 replication is 1, for $n = 30$.

	$n = 20$				$n = 30$				$n = 50$			
Method	Bias	TSD†	ESD	CP	Bias	TSD†	ESD	CP	Bias	TSD†	ESD	CP
						$\beta = 0.5$						
MCL*	0.026	0.772	0.811	0.972	0.019	0.608	0.629	0.958	0.020	0.456	0.469	0.960
JNF*	0.008	0.691	0.949	0.981	0.008	0.559	0.701	0.972	0.012	0.425	0.499	0.967
MDS*	0.008	0.664	0.686	0.944	0.006	0.548	0.560	0.942	0.011	0.426	0.435	0.949
MDS	0.008	0.664	0.686	0.944	0.006	0.548	0.560	0.942	0.011	0.426	0.435	0.949
						$\beta = 1.0$						
MCL*	0.078	0.759	0.803	0.967	0.056	0.622	0.629	0.961	0.038	0.460	0.471	0.956
JNF*	-0.038	0.654	0.935	0.977	-0.018	0.541	0.692	0.972	-0.005	0.435	0.496	0.971
MDS*	-0.014	0.665	0.675	0.932	-0.009	0.560	0.559	0.942	-0.002	0.436	0.438	0.940
MDS	-0.014	0.665	0.675	0.932	-0.009	0.560	0.559	0.942	-0.002	0.436	0.438	0.940
						$\beta = 1.5$						
MCL*	0.191	0.790	0.851	0.973	0.126	0.638	0.666	0.966	0.081	0.492	0.499	0.958
JNF*	-0.028	0.686	0.983	0.982	-0.007	0.576	0.731	0.982	0.010	0.471	0.522	0.963
MDS*	0.019	0.696	0.699	0.927	0.013	0.591	0.586	0.929	0.015	0.465	0.461	0.936
MDS	0.019	0.696	0.699	0.927	0.013	0.591	0.586	0.929	0.015	0.465	0.461	0.936
						$\beta = 2.0$						
MCL*	0.214	0.834	0.939	0.958	0.129	0.690	0.732	0.952	0.067	0.520	0.548	0.952
JNF*	-0.138	0.730	1.145	0.972	-0.059	0.634	0.816	0.956	-0.033	0.492	0.582	0.949
MDS*	-0.031	0.733	0.749	0.902	-0.026	0.631	0.635	0.914	-0.023	0.493	0.506	0.924
MDS	-0.028	0.734	0.751	0.902	-0.026	0.631	0.635	0.914	-0.023	0.493	0.506	0.924

Table 6. Results of the simulation study for two binary covariates and $M = 1$. MCL, JNF and MDS stand for the maximum conditional likelihood, jackknife, and the modified score estimators, respectively. TSD, ESD and CP represent the "true" standard error, estimated standard error and nominal 95% confidence interval coverage probability. †: approximation by the MAD method; *: estimate was calculated based on the datasets where both MCL and JNF exist out of 2000 replications; When $\beta_1 = \beta_2 = 0.5$ the number of divergent datasets out of 2000 replication are 208, and 10, $n = 30$ and 50 respectively, and when $\beta_1 = \beta_2 = 1$ the number of divergent datasets are 572, and 81, for $n = 30$, 50 respectively.

	$n = 30$				$n = 50$				$n = 100$			
Method	Bias	TSD†	ESD	CP	Bias	TSD†	ESD	CP	Bias	TSD†	ESD	CP
	$\beta_1 = 0.5, \beta_2 = 0.5$											
MCL*	0.014	0.621	0.642	0.980	0.016	0.484	0.486	0.959	0.005	0.338	0.331	0.950
	-0.030	0.625	0.684	0.985	0.043	0.522	0.524	0.963	0.017	0.348	0.355	0.963
JNF*	-0.082	0.510	0.759	0.997	-0.032	0.443	0.531	0.985	-0.014	0.324	0.344	0.966
	-0.131	0.495	0.819	0.998	-0.018	0.473	0.582	0.989	-0.006	0.334	0.371	0.980
MDS*	-0.045	0.551	0.556	0.965	-0.021	0.450	0.446	0.950	-0.012	0.325	0.318	0.946
	-0.088	0.549	0.586	0.970	-0.001	0.482	0.476	0.952	-0.003	0.336	0.340	0.962
MDS	-0.001	0.582	0.565	0.949	-0.019	0.451	0.447	0.950	-0.012	0.325	0.318	0.946
	0.007	0.605	0.596	0.940	0.006	0.482	0.477	0.949	-0.003	0.336	0.340	0.962
	$\beta_1 = 1.0, \beta_2 = 1.0$											
MCL*	-0.061	0.608	0.670	0.972	0.062	0.532	0.532	0.967	0.037	0.354	0.362	0.960
	-0.081	0.605	0.699	0.977	0.073	0.523	0.564	0.976	0.065	0.396	0.387	0.960
JNF*	-0.239	0.468	0.817	0.988	-0.051	0.468	0.605	0.975	-0.008	0.332	0.382	0.972
	-0.269	0.464	0.858	0.989	-0.055	0.450	0.649	0.985	0.014	0.373	0.411	0.980
MDS*	-0.169	0.527	0.566	0.945	-0.019	0.485	0.475	0.952	-0.002	0.336	0.343	0.951
	-0.192	0.515	0.586	0.96	-0.015	0.473	0.499	0.961	0.022	0.377	0.365	0.952
MDS	0.002	0.636	0.585	0.916	0.001	0.491	0.478	0.943	-0.002	0.336	0.343	0.951
	0.057	0.647	0.606	0.907	0.028	0.492	0.505	0.948	0.022	0.377	0.365	0.952

- The absolute bias of MCL estimator increases with β.

- The nominal 95% coverage probabilities for MCL estimator are either close to 0.95 or higher even though the MCL estimator has high bias. Further numerical investigation revealed that the MCL and its standard error are highly correlated. Therefore, for a large estimate of β, the standard error is also large, and thereby the confidence interval likely includes the true parameter.

- The jackknife method significantly reduces bias and the ESD of the estimate is always larger than its TSD.

- The MDS estimator has less absolute bias than that of the JNF method for nonzero β and its estimates were always finite for the simulation scenarios we considered. However, the MCL and JNF estimates were infinite in many occasions, specially for $n = 30$ or 50. For scenario 1 and when $n = 30$ and $\beta = 1.5$, about 50% datasets yield infinite MCL estimate (results not shown here).

- Overall, the variance of the MDS estimator is smaller than that of the other two estimators. We calculated ESD of the MDS estimator based on formula (3.3) which generally gives more accurate estimates of the true standard errors than that based on $\boldsymbol{I}(\widehat{\beta}_{\mathrm{R}})$. For example, in scenario 1 and when $\beta = 1$, the ESD based on formula (3.3) and $I(\widehat{\beta}_{\mathrm{R}})$ are 0.584 and 0.699 for $n = 30$, 0.482 and 0.582 for $n = 50$, and 0.349 and 0.369 for $n = 100$. The corresponding TSD are given in Table 4. Furthermore, the standard errors obtained from $I(\widehat{\beta}_{\mathrm{R}})$ are almost identical to the results obtained from $I(\widehat{\beta}_{\mathrm{MCL}})$ which are presented in the tables as the ESD of the MCL estimator.

- In some cases the MDS estimator has slightly lower coverage probabilities in its Wald-type confidence intervals. In unconditional logistic regression models, profile

likelihood intervals have been shown to have better coverage properties than Wald-type intervals (Heinze, 2006; Bull et al., 2007; and Heinze and Schemper, 2001a). Here we computed penalized-conditional-likelihood (PCL) based confidence intervals for the MDS estimator, and found that the corresponding coverage probability (CP) is close to 0.95. For example, in scenario 2, for $\beta = 2$ and $n = 20$ and 30 the PCL based coverage probabilities are 0.948 and 0.945, respectively. For small sample sizes, such as $n = 30$ and 50, the coverage probabilities based on the PCL confidence interval appear to be better than that based on a Wald-type confidence interval. Overall, the simulation results indicate that Wald-type intervals for MDS (no matter what variance estimate) do not cover well when MDS estimation is most desired.

- Additional simulation study (not shown here) indicates that all three methods yield almost unbiased estimates when the true value of β is zero.

For all three estimators in scenario 1 we also estimated the parameters by the method prescribed in Greenland (2000). For large values of β and for small sample size, Greenland's method reduces bias. For example, after deleting the datasets that MCL or JNF estimates are infinite, we found that when $\beta = 2$ (not shown) and sample size $n = 30$, the empirical bias due to Greenland and MCL estimators are $0.029\,(0.731)$ and $-0.725\,(0.427)$ based on 1166 remaining datasets and for $n = 50$ they are $0.131\,(0.668)$ and $-0.094\,(0.487)$, based on 1673 remaining datasets, respectively. The quantity in the parentheses represents the empirical standard error of the estimate, and it appears that the variance of the Greenland method is slightly larger than that of the MCL estimator. However, we did not find any appreciable differences between the Greenland's estimator and the MCL estimator for the sample sizes we considered in the simulation when β is 0.5, 1 and 1.5 (not shown).

In summary, the limited simulation study along with the data example below has illustrated the usefulness of the MDS estimator compared to the existing bias correction approaches in matched case-control studies.

3.5 An Analysis of Low-Birth-Weight Data

In order to illustrate the MDS method for $M = 1$ and $M > 1$ we considered the 1:3 matched low-birth-weight dataset discussed in Chapter I. Among several covariates, we focused on two covariates, the mother's smoking status (SMOKE) during the pregnancy and presence of previous preterm delivery (PTD).

First we considered 1:3 matching, and analyzed the data by the conditional logistic regression method, the jackknife method, and the modified score approach. The results are presented in the top part of Table 7. It is seen that at level 5% all three estimators MCL, JNF and MDS are statistically significant for PTD, while there is no significant association between SMOKE and the risk of having low-birth-weight child.

For the example of using $M = 1$, we randomly picked one control out of 3 from each stratum of the dataset and formed a 1:1 matched case-control data. The results of the analyses of the 1:1 matched data are presented in the bottom part of Table 7. Here we also considered SMOKE and PTD as the two covariates. Note that in this scenario the JNF estimates are infinite. At level 5%, the MDS estimate of the log-odds ratio for PTD is significant while the corresponding MCL estimate is not. Both the MDS and MCL estimates of the log-odds ratio parameter for SMOKE turn out to be statistically insignificant at level 5%. The table also provides the p-values which are calculated based on the asymptotic Z-statistic.

Table 7. Results of the analysis of the 1:M matched case-control data on low-birth-weight study with two covariates, SMOKE and PTD. The JNF estimator does not exist when $M = 1$. "Estimate" and "SE" denote the estimate and its standard error for the parameters of interest.

		SMOKE			PTD		
M	Method	Estimate	SE	P-value	Estimate	SE	P-value
3	MCL	0.598	0.476	0.209	1.733	0.611	0.005
	JNF	0.587	0.520	0.259	1.597	0.705	0.023
	MDS	0.583	0.461	0.206	1.646	0.565	0.004
1	MCL	0.699	0.615	0.256	1.896	1.083	0.080
	MDS	0.636	0.547	0.245	1.542	0.740	0.037

3.6 Discussion

In this chapter we apply Firth's general approach to reduce the bias in the maximum conditional likelihood estimator for matched case-control studies. The MDS estimator is obtained as the solution of a modified conditional score equation. The numerical studies showed that the MDS approach not only reduces bias but also has less asymptotic variance than the MCL estimators. The MDS technique yields finite parameter estimates, and can be applied when MCL and JNF estimates are infinite. Another advantage of the MDS approach over the JNF method is that the computation is easy and much less time consuming. Furthermore, like other bias preventive approaches the MDS method can also handle multiple covariates simultaneously, and the covariates could be categorical, continuous, or a mixture of both types.

It appears possible to apply the MDS method to obtain MDS estimates of the log-odds ratio parameters when a covariate is partially missing in the dataset. In this situation the likelihood will be more complex. A main issue in that context is how to appropriately calculate the information matrix. This problem deserves further investigation and is a topic for future research.

For the conditional logistic regression analysis we used `clogit` function of the statistical software `R`, and for the MDS estimator we used the Newton-Raphson method. Since Heinze and Schemper (2001a) adopted Firth's approach to handle the issue of monotonicity in Cox's partial likelihood and conditional logistic regression is a special case of Cox's partial likelihood for a stratified survival design, it is conceivable that one could adopt the software due to Heinze and Schemper (2001a) to obtain the MDS estimates.

CHAPTER IV

TESTING ADEQUACY OF A FUNCTIONAL FORM OF A COVARIATE IN MATCHED CASE-CONTROL STUDIES

4.1 Introduction

Suppose that Y, $\boldsymbol{X}$, and Z are the binary disease variable, a $p \times 1$ vector of covariates, and a continuous covariate of our interest, respectively. For modeling the disease risk in terms of the observed covariates we usually assume that

$$\text{pr}(Y = 1|\boldsymbol{X}, Z) = H\{\beta_0 + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_1 + \omega(Z; \boldsymbol{\beta}_2)\}\,,$$

where $\omega(Z; \boldsymbol{\beta}_2)$ is a known function of Z with some unknown parameters $\boldsymbol{\beta}_2$. The concern is whether $\omega(Z; \boldsymbol{\beta}_2)$ adequately captures the effect of Z. This is a well known problem of model goodness-of-fit. For instance, usually dietary fat increases the risk of breast cancer. However, the rate of increase of the risk for unit change in the amount of fat intake is not the same for the entire range of fat intake. Rather, it is a U-shaped function (Goodwin et al., 2003). Consequently a simple linear logistic effect of fat-intake will not adequately explain the effect on the risk of breast cancer. Therefore, it is scientifically important to check whether the fitted model adequately captures the effect of the potential risk factors. In this chapter we will investigate this issue in the context of matched case-control studies.

For matched case-control studies, Pregibon (1985), Moolgavkar et al. (1984), Moolgavkar et al. (1985), and Hosmer and Lemeshow (1989) developed conditional maximum likelihood-based diagnostic for detecting outliers and influential subjects. Hosmer and Lemeshow (1985) presented both parametric approaches based on probabilities and non-parametric methods to evaluate the ability of the multiple logistic regression model to distinguish the cases from the controls. Bedrick and Hill (1996) developed exact conditional

methods for checking the fit of a logistic regression to individual matched sets in case-control studies.

Arbogast and Lin (2004) proposed graphical and numerical methods for checking the adequacy of the logistic regression model for matched case-control studies. More specifically, they proposed three different tests for checking the overall model adequacy, the link function, and the functional forms of covariates, respectively. Arbogast and Lin's methods are based on cumulative sum of residuals over the covariates or linear predictors. The asymptotic distributions of their test statistics follow Gaussian processes, and the p-values of the tests are obtained by simulating a large number of empirical realizations of the approximate limiting processes. As a result the methods are very time consuming, and the computational burden increases heavily with the sample size. In this chapter, we are particularly interested in developing a simple and effective test for checking functional forms of a covariate.

This work is motivated by breast cancer data obtained from the Surveillance, Epidemiology and End Results (SEER) program which is a premier source for cancer statistics in the United States (`www.seer.cancer.gov/statfacts/html/breast.html`). The program routinely collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 26 percent of the US population. The SEER data contain information about patients' demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status. The age at diagnosis is one of important factors related to breast cancer prevention and control interventions, and it is also an important index for identifying the potential patients and deciding proper treatments for cancer patients. Generally, the effects of different functional forms of age at diagnosis on survival chance vary, thus the adequacy of the fitting needs to be checked in the logistic function of the risk of death due to breast cancer.

We propose to use a generalized score test for checking the adequacy of the functional form of a covariate. Boos (1992) discussed generalized score tests in the context of generalized estimating equations. The following is a brief outline of our method. Basically we write out an alternative model such that the fitted model whose adequacy is being tested is a special case of the alternative model. We approximate the alternative model by a regression spline with a given set of knot points. To avoid over-fitting the regression spline is accompanied by a penalty term which results in a penalized conditional likelihood function. In order to check the adequacy of the functional form of the covariate in the fitted model, we test whether the coefficients not corresponding to the assumed functional form are zero. The resulting score test involves the penalty parameter which is determined following the technique given by Gray (1994). Furthermore the asymptotic distribution of our proposed test statistic follows a linear combination of chi-square random variables.

The model and notation are described in Section 4.2. Section 4.3 contains the details of the proposed method. Section 4.4 describes the general application of the score test. Section 4.5 is the derivation of asymptotic distribution of test statistic T_n under the null model. Section 4.6 gives the power of the test statistic under the local alternative Model. Section 4.7 contains a simulation study to investigate the usefulness of the proposed method. In Section 4.8 we provide a data analysis on breast cancer data obtained from the SEER Program.

Before concluding this section we would like to emphasize the novel points of this chapter. First, we propose an effective method to address one important issue of testing the adequacy of the effect of a covariate on the risk of a disease that was studied previously only by Arbogast and Lin (2004). Second, the simulation results indicate that the proposed score test has much better power compared to Arbogast and Lin's test procedure. Third, in terms of the computational time, our proposed method is much faster than their simulation based test procedure.

4.2 Model and Notations

Suppose we have a 1:M_i (≥ 1) matched case-control data with n strata. Let Y_{ij} take on one or zero according as the j^{th} subject in the i^{th} matched set is a case or control respectively, where $i = 1, \cdots, n$, $j = 1, \cdots, M_i + 1$. Let $\boldsymbol{X}_{ij} = (X_{ij1}, \cdots, X_{ijp})^{\mathrm{T}}$ be a $p \times 1$ vector of covariates and Z is the covariate of interest or a suitable given transformation of the covariate. Let $\boldsymbol{S}_i$ be the set of covariates which are used for matching purposes in the i^{th} stratum. Define $\alpha_i(\boldsymbol{S}_i)$ as an arbitrary function which confers the effect of the i^{th} stratum on the risk of the disease. Under the null hypothesis that the effect of Z is linear, the disease risk model is

$$\text{pr}(Y_{ij} = 1|\boldsymbol{S}_i, \boldsymbol{X}_{ij}, Z_{ij}) = H\{\alpha_i(\boldsymbol{S}_i) + \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + \omega(Z_{ij}; \boldsymbol{\beta}_2)\}\,, \tag{4.1}$$

where $\boldsymbol{\beta}_1 = (\beta_{11}, \cdots, \beta_{1p})^{\mathrm{T}}$ is the vector of log-odds ratio parameters for the covariate $\boldsymbol{X}$ and $\boldsymbol{\beta}_2$ is a parameter or a vector of parameters related to the covariate Z.

The alternative model for the disease risk is assumed to be

$$\text{pr}(Y_{ij} = 1|\boldsymbol{S}_i, \boldsymbol{X}_{ij}, Z_{ij}) = H\{\alpha_i(\boldsymbol{S}_i) + \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + g(Z_{ij})\}\,, \tag{4.2}$$

where $g(\cdot)$ is an unknown smooth function with $m - 1$ continuous derivatives and square integrable m^{th} derivative on a compact set. We will first develop a method to test model (4.1) with $\omega(Z_{ij}; \beta_2) = Z_{ij}\beta_2$ against model (4.2). We then extend method to a more general case whose $\omega(Z_{ij}; \boldsymbol{\beta}_2)$ takes an arbitrary form.

4.3 Score Test Methodology

4.3.1 Derivation of the Test Statistic

Following Eubank (1988, p. 355) we will approximate the nonparametric function $g(\cdot)$ by a regression spline. Suppose that there are K known knot points $\kappa_1 < \cdots < \kappa_K$, such that $a < \kappa_1 < \cdots < \kappa_K < b$, where a and b are the minimum and maximum of the variable Z observed in the dataset. Define

$$\boldsymbol{C}_{1m}(Z) \equiv (Z, Z^2, \cdots, Z^m)^{\mathrm{T}}$$

and

$$\boldsymbol{C}_{2m}(Z) \equiv \{(Z-\kappa_1)_+^m, \cdots, (Z-\kappa_K)_+^m\}^{\mathrm{T}}$$

where $u_+ \equiv \max\{u, 0\}$.

Let $\boldsymbol{\gamma}_1 = (\gamma_{11}, \cdots, \gamma_{1m})^{\mathrm{T}}$ and $\boldsymbol{\gamma}_2 = (\gamma_{21}, \cdots, \gamma_{2K})^{\mathrm{T}}$. Now, $g(Z)$ can be parameterized by an m^{th} order regression spline function $\boldsymbol{C}_{1m}^{\mathrm{T}}(Z)\boldsymbol{\gamma}_1 + \boldsymbol{C}_{2m}^{\mathrm{T}}(Z)\boldsymbol{\gamma}_2$. Thus model (4.2) can be approximated by

$$\mathrm{pr}(Y_{ij} = 1 | \boldsymbol{S}_i, \boldsymbol{X}_{ij}, Z_{ij}) = H\{\alpha_i(\boldsymbol{S}_i) + \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{C}_{1m}^{\mathrm{T}}(Z_{ij})\boldsymbol{\gamma}_1 + \boldsymbol{C}_{2m}^{\mathrm{T}}(Z_{ij})\boldsymbol{\gamma}_2\}. \quad (4.3)$$

Kim et al. (2003) proposed a Bayesian method of estimation of model (4.3). However, they did not consider the problem of model adequacy. Define $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\gamma}_1^{\mathrm{T}}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$. Then the conditional log-likelihood function which is commonly used in matched case-control studies is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{M_i+1} Y_{ij} \log\{p_{ij}(\boldsymbol{\theta})\},$$

where

$$p_{ij}(\boldsymbol{\theta}) = \frac{\exp\{\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{C}_{1m}^{\mathrm{T}}(Z_{ij})\boldsymbol{\gamma}_1 + \boldsymbol{C}_{2m}^{\mathrm{T}}(Z_{ij})\boldsymbol{\gamma}_2\}}{\sum_{k=1}^{M_i+1} \exp\{\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{C}_{1m}^{\mathrm{T}}(Z_{ik})\boldsymbol{\gamma}_1 + \boldsymbol{C}_{2m}^{\mathrm{T}}(Z_{ik})\boldsymbol{\gamma}_2\}},$$

representing the conditional probability that the j^{th} subject is a case given that there is one case in the i^{th} stratum under model (4.3). To reduce the effect of over-fitting due to too many parameters, parameters are estimated by maximizing the penalized conditional log-likelihood function

$$l_p(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) - \frac{n\eta}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\theta}\,,$$

where η is a smoothing parameter that controls the degree of smoothing used, $\boldsymbol{D}$ is a $(p+m+K)\times(p+m+K)$ diagonal matrix with the first $p+m$ diagonal elements being zero followed by the last K diagonal elements all equal to one (Kim et al., 2003), i.e., $\boldsymbol{D} = \mathrm{diag}(\mathbf{0}_{p+1},\ \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}$ be the matrix obtained from $\boldsymbol{D}$ by deleting the first $p+1$ rows and columns.

Checking adequacy of model (4.1) is equivalent to test

$$H_0 : \gamma_{12} = \cdots \gamma_{1m} = 0 \ \text{ and } \boldsymbol{\gamma}_2 = \mathbf{0}\,.$$

Observe that under H_0 the disease risk model (4.3) reduces to (4.1).

Let $\boldsymbol{V}_{ij} \equiv (\boldsymbol{X}_{ij}^{\mathrm{T}}, \boldsymbol{C}_{1m}^{\mathrm{T}}(Z_{ij}), \boldsymbol{C}_{2m}^{\mathrm{T}}(Z_{ij}))^{\mathrm{T}}$. Under the disease risk model (4.3), the score function conditional on $\boldsymbol{S}$, $\boldsymbol{X}$, and Z becomes

$$\boldsymbol{U}_p(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{M_i+1}\{Y_{ij} - p_{ij}(\boldsymbol{\theta})\}\boldsymbol{V}_{ij} - n\eta\boldsymbol{D}\boldsymbol{\theta}$$

and the corresponding conditional information matrix is

$$\boldsymbol{I}_p(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta}) + n\eta\boldsymbol{D}\,,$$

where

$$\boldsymbol{I}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{M_i+1} p_{ij}(\boldsymbol{\theta})(\boldsymbol{V}_{ij} - \bar{\boldsymbol{V}}_{i\cdot})(\boldsymbol{V}_{ij} - \bar{\boldsymbol{V}}_{i\cdot})^{\mathrm{T}}$$

is the information matrix without the penalty term and $\bar{\boldsymbol{V}}_{i\cdot} \equiv \sum_{j=1}^{M_i+1} p_{ij}(\boldsymbol{\theta})\boldsymbol{V}_{ij}$.

Let

$$\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{M_i+1}\{Y_{ij} - p_{ij}(\boldsymbol{\theta})\}\boldsymbol{V}_{ij}\,,$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^{\mathrm{T}}, \gamma_{11})^{\mathrm{T}}$ and $\boldsymbol{\theta}_2 = (\gamma_{12}, \cdots, \gamma_{1m}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$. Now we partition functions

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{U}_1(\boldsymbol{\theta}) \\ \boldsymbol{U}_2(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2} \end{pmatrix}\,, \quad \boldsymbol{U}_p(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{U}_{p1}(\boldsymbol{\theta}) \\ \boldsymbol{U}_{p2}(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \frac{\partial l_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \\ \frac{\partial l_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2} \end{pmatrix}\,,$$

and the information matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) \;=\; \begin{pmatrix} \boldsymbol{I}_{11}(\boldsymbol{\theta}) & \boldsymbol{I}_{12}(\boldsymbol{\theta}) \\ \boldsymbol{I}_{21}(\boldsymbol{\theta}) & \boldsymbol{I}_{22}(\boldsymbol{\theta}) \end{pmatrix}\,.$$

Conditional on each stratum and given covariates $\boldsymbol{X}$ and Z, $\boldsymbol{I}_{11}(\boldsymbol{\theta}) = -E\{\partial \boldsymbol{U}_1(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_1^{\mathrm{T}}\}$ $= -\partial \boldsymbol{U}_1(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_1^{\mathrm{T}}$, $\boldsymbol{I}_{21}(\boldsymbol{\theta}) = -E\{\partial \boldsymbol{U}_2(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_1^{\mathrm{T}}\} = -\partial \boldsymbol{U}_2(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_1^{\mathrm{T}}$, $\boldsymbol{I}_{12}(\boldsymbol{\theta}) = \boldsymbol{I}_{21}^{\mathrm{T}}(\boldsymbol{\theta})$ and $\boldsymbol{I}_{22}(\boldsymbol{\theta}) = -E\{\partial \boldsymbol{U}_2(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_2^{\mathrm{T}}\} = -\partial \boldsymbol{U}_2(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_2^{\mathrm{T}}$.

Let $\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\theta}}_{01}^{\mathrm{T}}, \boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}}$ be the estimate of $\boldsymbol{\theta}$ under the null hypothesis H_0 which can be obtained by a simple conditional logistic regression analysis of model (4.1) with $\omega(Z_{ij}, \beta_2) = Z_{ij}\beta_2$. Following Boos's (1992) approach, expand the score function at the true parameter $\boldsymbol{\theta}$, i.e.,

$$\begin{aligned}
n^{-1/2}\boldsymbol{U}_{p1}(\hat{\boldsymbol{\theta}}_0) &= n^{-1/2}\boldsymbol{U}_{p1}(\boldsymbol{\theta}) + n^{-1/2}E\left(\frac{\partial \boldsymbol{U}_{p1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1^{\mathrm{T}}}\right)(\hat{\boldsymbol{\theta}}_{01} - \boldsymbol{\theta}_1) + \boldsymbol{O}_p(n^{-1/2}) \\
&= n^{-1/2}\boldsymbol{U}_{p1}(\boldsymbol{\theta}) + n^{-1/2}\boldsymbol{I}_{11}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{01} - \boldsymbol{\theta}_1) + \boldsymbol{O}_p(n^{-1/2}) \\
&= \boldsymbol{0}\,, \\
n^{-1/2}\boldsymbol{U}_{p2}(\hat{\boldsymbol{\theta}}_0) &= n^{-1/2}\boldsymbol{U}_{p2}(\boldsymbol{\theta}) + n^{-1/2}E\left(\frac{\partial \boldsymbol{U}_{p2}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1^{\mathrm{T}}}\right)(\hat{\boldsymbol{\theta}}_{01} - \boldsymbol{\theta}_1) + \boldsymbol{O}_p(n^{-1/2}) \\
&= n^{-1/2}\boldsymbol{U}_{p2}(\boldsymbol{\theta}) + n^{-1/2}\boldsymbol{I}_{21}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{01} - \boldsymbol{\theta}_1) + \boldsymbol{O}_p(n^{-1/2})\,.
\end{aligned}$$

Combining these two equations, then we have

$$\begin{aligned} n^{-1/2}\boldsymbol{U}_{p2}(\hat{\boldsymbol{\theta}}_0) &= n^{-1/2}\boldsymbol{U}_{p2}(\boldsymbol{\theta}) - n^{-1/2}\boldsymbol{I}_{21}(\boldsymbol{\theta})\boldsymbol{I}_{11}(\boldsymbol{\theta})^{-1}\boldsymbol{U}_{p1}(\boldsymbol{\theta}) + \boldsymbol{O}_p(n^{-1/2}) \\ &= n^{-1/2}\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{U}_p(\boldsymbol{\theta}) + \boldsymbol{O}_p(n^{-1/2})\,, \end{aligned}$$

where $\boldsymbol{A}(\boldsymbol{\theta}) = \left(-\boldsymbol{I}_{21}\boldsymbol{I}_{11}^{-1}, \boldsymbol{I}_{m-1+K}\right)$, an $(m-1+K)\times(p+m+K)$ matrix, and $\boldsymbol{I}_{m-1+K}$ is a $m-1+K$ identity matrix.

Notice that under H_0, $\boldsymbol{U}_p(\boldsymbol{\theta}) = \boldsymbol{U}(\boldsymbol{\theta})$. It implies that $E\{n^{-1/2}\boldsymbol{U}_2(\hat{\boldsymbol{\theta}}_0)\} = \boldsymbol{O}(n^{-1/2})$. Furthermore we have $\mathrm{var}\{\boldsymbol{U}_p(\boldsymbol{\theta})\} = \boldsymbol{I}_p(\boldsymbol{\theta})$, and

$$\begin{aligned} & \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{I}_p(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta})^{\mathrm{T}} \\ = \; & \left(-\boldsymbol{I}_{22}(\boldsymbol{\theta})\boldsymbol{I}_{11}^{-1}(\boldsymbol{\theta}), \boldsymbol{I}_{m-1+K}\right) \begin{pmatrix} \boldsymbol{I}_{11}(\boldsymbol{\theta}) & \boldsymbol{I}_{12}(\boldsymbol{\theta}) \\ \boldsymbol{I}_{21}(\boldsymbol{\theta}) & \boldsymbol{I}_{22}(\boldsymbol{\theta}) + n\eta\boldsymbol{\Lambda} \end{pmatrix} \begin{pmatrix} -\boldsymbol{I}_{11}^{-1}(\boldsymbol{\theta})\boldsymbol{I}_{12}(\boldsymbol{\theta}) \\ \boldsymbol{I}_{m-1+K} \end{pmatrix} \\ = \; & \boldsymbol{I}_{22}(\boldsymbol{\theta}) - \boldsymbol{I}_{21}(\boldsymbol{\theta})\boldsymbol{I}_{11}^{-1}(\boldsymbol{\theta})\boldsymbol{I}_{12}(\boldsymbol{\theta}) + n\eta\boldsymbol{\Lambda}\,. \end{aligned}$$

Thus the score test statistic can be given by

$$T_n = \boldsymbol{U}_2^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_0)\{\boldsymbol{I}_{22}(\hat{\boldsymbol{\theta}}_0) - \boldsymbol{I}_{21}(\hat{\boldsymbol{\theta}}_0)\boldsymbol{I}_{11}^{-1}(\hat{\boldsymbol{\theta}}_0)\boldsymbol{I}_{12}(\hat{\boldsymbol{\theta}}_0) + n\eta\boldsymbol{\Lambda}\}^{-1}\boldsymbol{U}_2(\hat{\boldsymbol{\theta}}_0)\,. \tag{4.4}$$

4.3.2 Asymptotic Distribution of the Test Statistic under the Null Model

Suppose that the number of knot points is fixed as sample size goes to ∞, and the usual conditions are satisfied so that the standard asymptotic expansion holds for unpenalized conditional likelihood. Then the test statistic T_n is asymptotically distributed as $\sum_j \delta_j G_j^2$ under the null hypothesis, where $\delta_j s$ are the positive eigenvalues of the matrix $\lim_{n\to\infty}(\boldsymbol{I}_{22} - \boldsymbol{I}_{21}\boldsymbol{I}_{11}^{-1}\boldsymbol{I}_{12})(\boldsymbol{I}_{22} - \boldsymbol{I}_{21}\boldsymbol{I}_{11}^{-1}\boldsymbol{I}_{12} + n\eta\boldsymbol{\Lambda})^{-1}$ and $G_j s$ are independent standard normal variables. A simple proof is given in section 4.5. Thus for testing the hypothesis, the remaining main issue is how to obtain the distribution of the linear combination of several χ_1^2 distributions effectively. Here we will approximate the distribution by the Satterthwaite

method. The basic idea is to find an effective degrees of freedom ν such that $\nu T_n/E(T_n)$ approximately follows the χ^2_ν distribution. Now, using the fact that $E(T_n) = \sum_j \delta_j$ and $\mathrm{var}(T_n) = 2\sum_j \delta_j$, then

$$\mathrm{var}\left\{\nu\frac{T_n}{E(T_n)}\right\} = \nu^2\mathrm{var}\left\{\frac{T_n}{E^2(T_n)}\right\} = 2\nu^2\frac{\sum_j \delta_j^2}{(\sum_j \delta_j)^2}\,.$$

Equating $\mathrm{var}\{\nu T_n/E(T_n)\} = \mathrm{var}(\chi^2_\nu) = 2\nu$, we obtain $\nu = (\sum \delta_j)^2/\sum {\delta_j}^2$. Therefore, we consider a scaled version of T_n,

$$Q_n \equiv T_n(\sum_j \delta_j)/(\sum_j \delta_j^2)$$

as our final test statistic and reject the null hypothesis at 5% level of significance if $Q_n > \chi^2_{\nu,0.05}$, the 95^{th} quantile of the χ^2_ν distribution.

The power of the test statistic T_n can be calculated theoretically. The details are given in section 4.6.

4.3.3 Choice of the Penalty Parameter and the Knot Points

Now some remarks are in order about the proposed test statistic. First of all, for the calculation of the test statistic, we only need to estimate the parameters of the null model (4.1). However, statistic T_n involves the penalty parameter η through δ_j^2; and η and the degrees of freedom ν are interrelated. The range of η is $(0, \infty)$, and a larger value of η leads to a smoother $g(Z)$. On the other hand, for finite sample sizes, the data may not capture the nonparametric effect very accurately, so one needs to use smaller degrees of freedom to test the null hypothesis. Thus the range of possible degrees of freedom is much shorter than the range of the penalty parameter. Therefore, following Gray (1994) we will specify the degrees of freedom and then determine the penalty parameter η. For moderate to large sample sizes we have tried different degrees of freedom (df) from 2 to 4. All the degrees

of freedom appear to be reasonable. To be specific, we recommend taking 3 as degrees of freedom. Furthermore, for given degrees of freedom we used 2, 3, 5 and 10 knot points. We have chosen quantiles of the observed covariate Z that we are interested in as the knot points. Specifically, when the number of knots is 2, 3 or 5 we choose equal spaced quantiles between 30^{th} and 70^{th} percentiles, and when the number of knot points is 10 we choose equal spaced quantiles between 20^{th} and 80^{th} percentiles.

4.4 Generalization for Arbitrary Known Form of $\omega(Z;\boldsymbol{\beta}_2)$

We assume that $\omega(Z;\boldsymbol{\beta}_2)$ is not a spline which is a practical assumption when $\omega(Z;\boldsymbol{\beta}_2)$ is assumed to be known to a practitioner. Without loss of generality we assume

$$\omega(Z;\boldsymbol{\beta}_2) = \sum_{k=1}^{m_0} \rho_k(\boldsymbol{\beta}_2) Z^k + \omega_2(Z;\boldsymbol{\beta}_2)\,,$$

where $m_0 \leq m$ and $\omega_2(Z;\boldsymbol{\beta}_2)$ is a non-polynomial function. If $\rho_k(\boldsymbol{\beta}_2) \equiv 0$ for all $1 \leq k \leq m_0$, define $\boldsymbol{C}^-_{1m}(Z) = \boldsymbol{C}_{1m}(Z)$, otherwise let $\boldsymbol{C}^-_{1m}(Z)$ be the vector consisting of remaining components of $\boldsymbol{C}_{1m}(Z)$ after deleting all Z^k corresponding to $\rho_k(\boldsymbol{\beta}_2) \neq 0$ for $1 \leq k \leq m_0$, and correspondingly we define $\boldsymbol{\gamma}_1^-$. Now, the alternative model which covers the assumed parametric model as a special case can be written as

$$\begin{aligned} \text{pr}(Y_{ij}=1|\boldsymbol{S}_i,\boldsymbol{X}_{ij},Z_{ij}) &= H\{\alpha_i(\boldsymbol{S}_i) + \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + \omega(Z_{ij};\boldsymbol{\beta}_2) \\ &+ \boldsymbol{C}_{1m}^{-\mathrm{T}}(Z_{ij})\boldsymbol{\gamma}_1^- + \boldsymbol{C}_{2m}^{\mathrm{T}}(Z_{ij})\boldsymbol{\gamma}_2\}\,. \end{aligned} \tag{4.5}$$

Thus for testing adequacy of the known form $\omega(Z;\boldsymbol{\beta}_2)$ we need test

$$H_0 : \boldsymbol{\gamma}_1^- = \mathbf{0} \text{ and } \boldsymbol{\gamma}_2 = \mathbf{0}\,,$$

following steps similar to those described in Section 4.3.

Specifically, for $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \boldsymbol{\gamma}^{-^{\mathrm{T}}}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$, here

$$\begin{aligned} p_{ij}(\boldsymbol{\theta}) &= \frac{\exp\{\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + \omega(Z;\boldsymbol{\beta}_2) + \boldsymbol{C}_{1m}^{\mathrm{T}}(Z_{ij})\boldsymbol{\gamma}_1 + \boldsymbol{C}_{2m}^{\mathrm{T}}(Z_{ij})\boldsymbol{\gamma}_2\}}{\sum_{k=1}^{M_i+1}\exp\{\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta}_1 + \omega(Z;\boldsymbol{\beta}_2) + \boldsymbol{C}_{1m}^{\mathrm{T}}(Z_{ik})\boldsymbol{\gamma}_1 + \boldsymbol{C}_{2m}^{\mathrm{T}}(Z_{ik})\boldsymbol{\gamma}_2\}}\,, \\ \boldsymbol{U}(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{j=1}^{M_i+1}\{Y_{ij} - p_{ij}(\boldsymbol{\theta})\}\boldsymbol{V}_{ij}\,, \end{aligned}$$

and

$$\frac{\partial \boldsymbol{U}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = \sum_{i=1}^{n}\left\{\sum_{j=1}^{M_i+1}(Y_{ij} - p_{ij})\tilde{D} - \sum_{j=1}^{M_i+1} p_{ij}(\boldsymbol{\theta})(\boldsymbol{V}_{ij} - \bar{\boldsymbol{V}}_{i\cdot})(\boldsymbol{V}_{ij} - \bar{\boldsymbol{V}}_{i\cdot})^{\mathrm{T}}\right\},$$

where $V_{ij} = \left(\boldsymbol{X}_{ij}^{\mathrm{T}}, \partial^{\mathrm{T}}\omega(Z;\boldsymbol{\beta}_2)/\partial\boldsymbol{\beta}_2, \boldsymbol{C}_{1m}^{\mathrm{T}}(Z_{ij}), \boldsymbol{C}_{2m}^{\mathrm{T}}(Z_{ij})\right)^{\mathrm{T}}$, $\bar{\boldsymbol{V}}_{i\cdot} \equiv \sum_{j=1}^{M_i+1} p_{ij}(\boldsymbol{\theta})\boldsymbol{V}_{ij}$, and $\tilde{D} = \mathrm{diag}(\mathbf{0}_p, \boldsymbol{I}_{|\boldsymbol{\beta}_2|}, \mathbf{0}_{|\boldsymbol{\gamma}_1^-|}, \mathbf{0}_{|\boldsymbol{\gamma}_2|})$. Here $\mathbf{0}_{|\boldsymbol{t}|}$ represent the square matrix whose dimension is the same as the vector $\boldsymbol{t}$, and $\boldsymbol{I}_{|\boldsymbol{\beta}_2|}$ is an identity matrix. Thus, Under H_0,

$$\boldsymbol{I}(\boldsymbol{\theta}) = -E\left\{\frac{\partial \boldsymbol{U}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\} = \sum_{i=1}^{n}\sum_{j=1}^{M_i+1} p_{ij}(\boldsymbol{\theta})(\boldsymbol{V}_{ij} - \bar{\boldsymbol{V}}_{i\cdot})(\boldsymbol{V}_{ij} - \bar{\boldsymbol{V}}_{i\cdot})^{\mathrm{T}}\,.$$

For the general case considered here,

$$\boldsymbol{U}_2(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{M_i+1}\{Y_{ij} - p_{ij}(\boldsymbol{\theta})\}\boldsymbol{V}_{2ij}$$

with $\boldsymbol{V}_{2ij} = (C_{1m}^{\mathrm{T}}(Z_{ij}), C_{2m}^{\mathrm{T}}(Z_{ij}))^{\mathrm{T}}$.

4.5 Derivation of Asymptotic Distribution of Test Statistic T_n under the Null Model

We derive the asymptotic distribution of the statistic T_n under the null model (4.1). Let $\boldsymbol{V}_{ij,2} = \{\boldsymbol{C}_{1m}^{-^{\mathrm{T}}}(Z_{ij}), \boldsymbol{C}_2^{\mathrm{T}}(Z_{ij})\}$, and E and Cov be the expectation and covariance condi-

tional on $\boldsymbol{S}$, $\boldsymbol{X}$ and Z under the null model. Define

$$\begin{aligned}
\boldsymbol{P} &\equiv \lim_{n\to\infty}\frac{1}{n}E\{\boldsymbol{I}_{22}(\hat{\boldsymbol{\theta}}_0)-\boldsymbol{I}_{21}(\hat{\boldsymbol{\theta}}_0)\boldsymbol{I}_{11}^{-1}(\hat{\boldsymbol{\theta}}_0)\boldsymbol{I}_{12}(\hat{\boldsymbol{\theta}}_0)\}+\eta\boldsymbol{\Lambda},\\
\boldsymbol{W}_2 &\equiv \lim_{n\to\infty}\frac{1}{n}\mathrm{Cov}\left\{\boldsymbol{U}_2(\hat{\boldsymbol{\theta}}_0)\right\}\\
&= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M_i+1}p_{ij}(\boldsymbol{\theta})\left\{\boldsymbol{V}_{ij,2}-\bar{\boldsymbol{V}}_{i\cdot,2}\right\}\left\{\boldsymbol{V}_{ij,2}-\bar{\boldsymbol{V}}_{i\cdot,2}\right\}^{\mathrm{T}},
\end{aligned}$$

where $p_{ij}(\boldsymbol{\theta})=\exp(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1)/\sum_{k=1}^{M_i+1}\exp(\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta}_1)$, and $\bar{\boldsymbol{V}}_{i\cdot,2}=\sum_{j=1}^{M_i+1}p_{ij}(\boldsymbol{\theta})\boldsymbol{V}_{ij,2}$. Let $\boldsymbol{J}\equiv n^{-1/2}\boldsymbol{P}^{-1/2}\boldsymbol{U}_2(\hat{\boldsymbol{\theta}}_0)$. Then $E(\boldsymbol{J})=0$ and $\boldsymbol{W}\equiv\mathrm{Cov}(\boldsymbol{J})=\boldsymbol{P}^{-1/2}\boldsymbol{W}_2\boldsymbol{P}^{-1/2}$. Note that $\boldsymbol{J}\to\mathrm{Normal}(\boldsymbol{0},\boldsymbol{W})$ in distribution. Thus $\boldsymbol{W}^{-1/2}\boldsymbol{J}$ asymptotically follows a normal distribution with mean $\boldsymbol{0}$ and $(m+K-1)\times(m+K-1)$ identity covariance matrix $\boldsymbol{I}_{m+K-1}$.

Since $\boldsymbol{W}$ is a positive definite matrix, by the factorization theorem we can write $\boldsymbol{W}=\boldsymbol{R}\boldsymbol{\Delta}\boldsymbol{R}^{\mathrm{T}}$, where $\boldsymbol{R}$ is the matrix of the orthogonal eigenvectors and $\boldsymbol{\Delta}$ is the diagonal matrix with diagonal eigenvalues λ_j of $\boldsymbol{W}$. Then $\boldsymbol{\Gamma}=\boldsymbol{R}^{\mathrm{T}}\boldsymbol{W}^{-1/2}\boldsymbol{J}\to\mathrm{Normal}(\boldsymbol{0},\boldsymbol{I}_{m+K-1})$ in distribution. Therefore, $T_n=\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}=\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{\Gamma}\to\sum_j\lambda_jB_j^2$ in distribution as $n\to\infty$. Here B_j^2s are independent and each of which follows a centered χ^2 distribution with 1 degree of freedom.

4.6 Power Consideration under the Local Alternative Model

We derive the asymptotic distribution of the statistic T_n under the local alternative model (4.2) with the departure of $g(Z)$ from its null mode (4.1) in the order of $O(n^{-1/2})$. Let $\boldsymbol{V}_{ij,2}=\{\boldsymbol{C}_{1m}^{-\mathrm{T}}(Z_{ij}),\boldsymbol{C}_2^{\mathrm{T}}(Z_{ij})\}$, and E_* and Cov_* be the expectation and covariance condi-

tional on $\boldsymbol{S}$, $\boldsymbol{X}$ and Z under the true model. Define

$$\begin{aligned}
\boldsymbol{P} &\equiv \lim_{n\to\infty}\frac{1}{n}E_*\{\boldsymbol{I}_{22}(\hat{\boldsymbol{\theta}}_0)-\boldsymbol{I}_{21}(\hat{\boldsymbol{\theta}}_0)\boldsymbol{I}_{11}^{-1}(\hat{\boldsymbol{\theta}}_0)\boldsymbol{I}_{12}(\hat{\boldsymbol{\theta}}_0)\}+\eta\boldsymbol{\Lambda},\\
\boldsymbol{\mu}_2 &\equiv \lim_{n\to\infty}\frac{1}{\sqrt{n}}E_*\{\boldsymbol{U}_2(\hat{\boldsymbol{\theta}}_0)\}=\lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{M_i+1}\left\{p_{ij}^*(\boldsymbol{\theta})-E_*\left(p_{ij}^{**}(\hat{\boldsymbol{\theta}}_0)\right)\right\}\boldsymbol{V}_{ij,2},\\
\boldsymbol{W}_2 &\equiv \lim_{n\to\infty}\frac{1}{n}\mathrm{Cov}_*\left\{\boldsymbol{U}_2(\hat{\boldsymbol{\theta}}_0)\right\}\\
&= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M_i+1}p_{ij}^*(\boldsymbol{\theta})\left\{\boldsymbol{V}_{ij,2}-\bar{\boldsymbol{V}}_{i\cdot,2}\right\}\left\{\boldsymbol{V}_{ij,2}-\bar{\boldsymbol{V}}_{i\cdot,2}\right\}^{\mathrm{T}},
\end{aligned}$$

where $p_{ij}^*(\boldsymbol{\theta})=\exp\{\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1+g(Z_{ij})\}/\sum_{k=1}^{M_i+1}\exp\{\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta}_1+g(Z_{ij})\}$, $p_{ij}^{**}(\boldsymbol{\theta})=\exp(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1)/\sum_{k=1}^{M_i+1}\exp(\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta}_1)$, and $\bar{\boldsymbol{V}}_{i\cdot,2}=\sum_{j=1}^{M_i+1}p_{ij}^*(\boldsymbol{\theta})\boldsymbol{V}_{ij,2}$.

Let $\boldsymbol{J}\equiv n^{-1/2}\boldsymbol{P}^{-1/2}\boldsymbol{U}_2(\hat{\boldsymbol{\theta}}_0)$. Then $E_*(\boldsymbol{J})=\boldsymbol{P}^{-1/2}\boldsymbol{\mu}_2$ and $\boldsymbol{W}\equiv\mathrm{Cov}_*(\boldsymbol{J})=\boldsymbol{P}^{-1/2}\boldsymbol{W}_2\boldsymbol{P}^{-1/2}$. Note that $\boldsymbol{J}\to\mathrm{Normal}(\boldsymbol{P}^{-1/2}\boldsymbol{\mu}_2,\boldsymbol{W})$ in distribution. Thus $\boldsymbol{W}^{-1/2}\boldsymbol{J}$ asymptotically follows a normal distribution with mean $\boldsymbol{W}^{-1/2}\boldsymbol{P}^{-1/2}\boldsymbol{\mu}_2$ and $(m+K-1)\times(m+K-1)$ identity covariance matrix $\boldsymbol{I}_{m+K-1}$.

Since $\boldsymbol{W}$ is a positive definite matrix, by the factorization theorem we can write $\boldsymbol{W}=\boldsymbol{R}\boldsymbol{\Delta}\boldsymbol{R}^{\mathrm{T}}$, where $\boldsymbol{R}$ is the matrix of the orthogonal eigenvectors and $\boldsymbol{\Delta}$ is the diagonal matrix with diagonal eigenvalues λ_j of $\boldsymbol{W}$. Denote $\boldsymbol{\tau}=\boldsymbol{R}^{\mathrm{T}}\boldsymbol{W}^{-1/2}\boldsymbol{P}^{-1/2}\boldsymbol{\mu}_2$. Then $\boldsymbol{\Gamma}=\boldsymbol{R}^{\mathrm{T}}\boldsymbol{W}^{-1/2}\boldsymbol{J}\to\mathrm{Normal}(\boldsymbol{\tau},\boldsymbol{I}_{m+K-1})$ in distribution. Therefore, $T_n=\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}=\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{\Gamma}\to\sum_j\lambda_jB_j^2$ in distribution as $n\to\infty$. Here B_j^2s are independent and each of which follows a non-central χ^2 distribution with 1 degree of freedom and noncentrality parameter τ_j^2, where τ_j is the j^{th} component of $\boldsymbol{\tau}$. Furthermore, by the Satterthwaite approximation, the power of the score test may be approximated by $\mathrm{pr}((\sum_j\delta_j^2)^{-1}\sum_j\delta_j\sum_j\lambda_jB_j^2>\chi^2_{\nu,0.05})$, where δ_js are eigenvalues of matrix $(\boldsymbol{P}-\eta\boldsymbol{\Lambda})\boldsymbol{P}^{-1}$.

4.7 A Simulation Study

In order to judge the performance of the proposed method we performed the following simulation study. We first generated a cohort data of size $16,000$ with variables S, X, Z and Y according to the following two simulation scenarios.

1. $S \sim \text{Normal}(0,1)$, $Z \sim \text{Normal}(S,1)$ and $Y \sim \text{Bernoulli}(p_y)$, where $p_y = H(\alpha_0 + 1.1S - 0.25Z + \beta Z^2)$; with $\alpha_0 = -2.4$ for $\beta = 0$, $\alpha_0 = -2.6$ for $\beta = 0.10$, and $\alpha = -3.0$ for $\beta = 0.25$.

2. $S \sim \text{Normal}(0,1)$, $Z \sim \text{Normal}(S,1)$, $X \sim \text{Bernoulli}(p_z)$ with $p_z = H(-0.5 + S)$ and $Y \sim \text{Bernoulli}(p_y)$, where $p_y = H(\alpha_0 + 1.1S + 0.5X - 0.25Z + \beta Z^2)$; with $\alpha_0 = -2.7$ for $\beta = 0$, $\alpha_0 = -2.9$ for $\beta = 0.10$, and $\alpha_0 = -3.3$ for $\beta = 0.25$.

The true parameters of the distribution of S and the log-odds ratio parameter for S in p_y were chosen based on the setting given in Arbogast and Lin (2004). We chose a value of α_0 so that the overall marginal disease prevalence is around 10%. From the cohort data we created 1:3 matched case-control data with n strata using S as the matching variable. For scenario 1 we considered only a single covariate, and for scenario 2 we considered two covariates. For both scenarios we considered three different sample sizes $n = 25$, 50, and 100. Note that for both the scenarios the true effect of Z on the disease risk is a quadratic when β equals 0.10 and 0.25. For both scenarios, the true effect of Z is linear when β equals zero. Thus $\beta = 0$ situation allows us to judge the level of the proposed test.

Our results are based on $N = 1000$ datasets, and for each scenario we calculated the power of our proposed test and that of Arbogast and Lin's (2004) approach. In our proposed score method, we approximated $g(\cdot)$ by a cubic regression spline, i.e., $m = 3$. Note that the limiting distribution of Arbogast and Lin's (2004) test G_2 follows a Gaussian process, for each dataset. The p-value calculation of the test is based on $B = 1000$ bootstrap samples.

Under scenario 1 with $n = 50$, $df = 2.5$, and $\beta = 0$, the computational time in minutes for Q_n and G_2 are 40.3 and 1594, respectively, and a ratio of about 1:39 when the program coded in R was run in a 2.6 GHZ Xenon processor.

The simulation results for scenarios 1 and 2 are presented in Tables 8 and 9, respectively. The performance of two methods can be summarizes as follows:

- These results indicate that the empirical levels of both penalized score test and the test of Arbogast and Lin (2004) are not significantly different from the nominal level 0.05.

- For a given sample size, the power of both tests increases with β. Furthermore, the power of the tests increases with the sample size as expected.

- The results also indicate that in most of the cases the number of knot points and degrees of freedom in the proposed score test do not affect the power of the test significantly for the fixed β at a given sample size. Only in a few cases, the power of the score test varies somewhat. For example, in scenario 2 (Table 9), when $n = 25$ and $\beta = 0.10$, the power is 0.213 and 0.215 for 5 and 10 knot points, respectively; while the power of the other cases is round 0.13.

- The results in both scenarios show that for $\beta = 0.10$ at all considered sample sizes and for $\beta = 0.25$ with sample size 25, the power of our proposed test is at least twice of that of Arbogast and Lin (2004).

Overall the power of the proposed test is much higher than that of Arbogast and Lin (2004). Importantly, the power of the proposed test is generally remarkably stable for changing degrees of freedom and number of knot points.

Table 8. Power comparison from simulation studies for a single covariate. Here G_2 represents the test statistic from Arbogast and Lin (2004). The levels of the tests are listed under $\beta = 0$ compared to the nominal level 0.05.

		Power of the Score Test											
n		$\beta = 0$				$\beta = 0.10$				$\beta = 0.25$			
		Number of Knots				Number of Knots				Number of Knots			
	df	2	3	5	10	2	3	5	10	2	3	5	10
	2	0.045	0.045	0.042	0.045	0.139	0.141	0.208	0.141	0.546	0.546	0.854	0.855
	2.5	0.050	0.050	0.049	0.049	0.151	0.151	0.153	0.153	0.553	0.552	0.552	0.547
25	3	0.054	0.054	0.057	0.055	0.154	0.150	0.149	0.153	0.535	0.532	0.528	0.533
	3.5	0.062	0.056	0.056	0.055	0.144	0.147	0.131	0.134	0.517	0.520	0.510	0.517
	4	0.056	0.056	0.053	0.053	0.120	0.134	0.135	0.140	0.453	0.516	0.492	0.501
	2	0.041	0.042	0.042	0.042	0.207	0.208	0.208	0.208	0.852	0.854	0.854	0.855
	2.5	0.046	0.046	0.046	0.046	0.216	0.216	0.216	0.214	0.845	0.845	0.844	0.846
50	3	0.047	0.052	0.053	0.052	0.208	0.207	0.204	0.207	0.831	0.829	0.823	0.829
	3.5	0.053	0.052	0.052	0.050	0.195	0.197	0.198	0.195	0.820	0.817	0.817	0.819
	4	0.053	0.052	0.054	0.055	0.168	0.197	0.184	0.188	0.786	0.817	0.805	0.813
	2	0.063	0.063	0.063	0.062	0.338	0.338	0.338	0.336	0.988	0.988	0.988	0.988
	2.5	0.058	0.058	0.058	0.058	0.338	0.339	0.339	0.339	0.990	0.990	0.989	0.990
100	3	0.061	0.062	0.061	0.061	0.331	0.330	0.330	0.331	0.987	0.986	0.986	0.986
	3.5	0.062	0.062	0.063	0.063	0.323	0.323	0.323	0.323	0.982	0.982	0.982	0.982
	4	0.064	0.062	0.062	0.063	0.294	0.310	0.310	0.323	0.976	0.981	0.981	0.982

Power of G_2

n	$\beta = 0$	$\beta = 0.10$	$\beta = 0.25$
25	0.047	0.062	0.210
50	0.048	0.097	0.608
100	0.051	0.167	0.790

Table 9. Power comparison from simulation studies for two covariates. Here G_2 represents the test statistics from Arbogast and Lin (2004). The levels of the tests are listed under $\beta = 0$ compared to the nominal level 0.05.

		Power of the Score Test											
n		$\beta = 0$				$\beta = 0.10$				$\beta = 0.25$			
		Number of Knots				Number of Knots				Number of Knots			
	df	2	3	5	10	2	3	5	10	2	3	5	10
	2	0.050	0.050	0.050	0.050	0.126	0.126	0.213	0.215	0.549	0.549	0.550	0.550
	2.5	0.054	0.055	0.055	0.055	0.139	0.139	0.138	0.139	0.529	0.530	0.528	0.530
25	3	0.056	0.055	0.056	0.056	0.140	0.136	0.133	0.137	0.552	0.546	0.540	0.548
	3.5	0.057	0.056	0.055	0.056	0.127	0.128	0.127	0.127	0.540	0.538	0.534	0.539
	4	0.056	0.056	0.055	0.055	0.123	0.125	0.126	0.132	0.492	0.492	0.519	0.522
	2	0.051	0.047	0.047	0.048	0.220	0.213	0.215	0.216	0.846	0.839	0.841	0.855
	2.5	0.047	0.047	0.047	0.047	0.234	0.223	0.223	0.214	0.823	0.825	0.825	0.846
50	3	0.050	0.048	0.047	0.048	0.215	0.216	0.216	0.215	0.835	0.829	0.829	0.832
	3.5	0.056	0.046	0.045	0.046	0.210	0.209	0.206	0.212	0.812	0.813	0.812	0.814
	4	0.041	0.041	0.041	0.044	0.179	0.194	0.189	0.204	0.769	0.769	0.769	0.801
	2	0.055	0.056	0.055	0.062	0.366	0.366	0.338	0.365	0.988	0.988	0.988	0.988
	2.5	0.054	0.053	0.053	0.053	0.358	0.357	0.357	0.358	0.984	0.984	0.984	0.984
100	3	0.051	0.050	0.053	0.051	0.348	0.346	0.346	0.346	0.988	0.987	0.987	0.987
	3.5	0.054	0.054	0.054	0.054	0.330	0.333	0.332	0.333	0.986	0.984	0.983	0.985
	4	0.052	0.055	0.055	0.053	0.279	0.314	0.311	0.319	0.975	0.981	0.981	0.979

Power of G_2			
n	$\beta = 0$	$\beta = 0.10$	$\beta = 0.25$
25	0.038	0.060	0.210
50	0.055	0.084	0.440
100	0.060	0.152	0.807

4.8 An Application to the SEER Breast Cancer Data

We now return to the SEER data discussed in the introduction. The data (1973-2003) from National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, were based on the November 2005 submission and were publicly available in April 2006. The follow-up cutoff date was December 31, 2003. For the purpose of our analysis we consider only white and black females aged 30 years or older from Connecticut registry and diagnosed for breast cancer between January, 1980 and December, 2000.

Since the grades of tumor play an important role in the survival time, we will use it as one of the matching variables. The grades of cancer were described as well differentiated (I), moderately differentiated (II), poorly differentiated (III), undifferentiated (IV), and others (all others except for the above four categories). In the SEER data estrogen receptor status was recorded as tumor marker 1 (Bedrosian et al., 2008). Estrogen receptor positive (ER+) means that estrogen is causing the tumor to grow, and that the cancer should respond well to hormone suppression treatments. If the estrogen receptor status is negative (ER-), then the tumor is not driven by estrogen, thus an effective treatment need to be determined. Our data analysis focused only on the black or white women whose tumor marker 1 is either 'Positive' or 'Negative' and grades ranges from I to IV. Thus the cohort contains $16,930$ women, of which $5,194$ died before December 31, 2003. The women who died will be treated as cases ($Y = 1$) and otherwise controls ($Y = 0$). First we randomly selected $n = 200$ woman who died on or before the end of the study (cases) from the cohort of 16,930 women, then for each randomly chosen case we selected M control women by matching the race and cancer grade's level. In this example we took $M = 3$. In order to avoid repetition of the same subjects, once a control was matched with a case, it was then removed from the cohort. The goal is to test whether the age at diagnosis of breast cancer has a linear effect on the survival of a patient.

Define a binary variable X corresponding to the tumor marker 1 as

$$X = \begin{cases} 1 & \text{if tumor marker 1 is positive} \\ 0 & \text{otherwise}\,. \end{cases}$$

The age at diagnosis was measured in years. In our matched case-control data the age at diagnosis ranges from 31 to 97 years, with a mean of 61.8 years and a standard deviation of 14.2 years. For the analysis purpose we define continuous variable

$$Z = \frac{\text{age at diagnosis} - 61.8}{14.2}\,.$$

First we fit a model

$$\text{pr}(Y = 1|\boldsymbol{S}, X, Z) = H(\beta_0(\boldsymbol{S}) + X\beta_1 + Z\beta_2)\,,$$

where $\boldsymbol{S}$ consists of two variables race and grade. The results are presented in Table 10. Here ER and SE stand for estrogen receptor (tumor marker 1) and the estimated standard error. The results show that ER+ has a significantly negative impact on the risk of survival

Table 10. Conditional logistic regression analysis of the 1:3 matched case-control data constructed from the SEER study.

Covariate	Estimate	SE	p-value
ER positive vs. negative	-0.427	0.205	0.037
Age at Diagnosis	0.856	0.105	< 0.001

which supports the fact that ER+ subjects respond well to their hormone suppression treatment as all the subjects were under some therapy or treatment after the diagnosis of the cancer. Age at diagnosis has a statistically significant effect on the survival.

Now we test if the linear effect of age at diagnosis adequately explains its association with the survival. The proposed test statistic Q_n with 3 degrees of freedom and 10 knot points is 10.025, which gives a p-value of 0.018. We also analyzed the data with degrees

of freedom 2, 2.5, 3, 3.5 and 4 with number of knot points 2, 3, 4, and 10. The test statistic Q_n varies little for different number of knot points. As is expected, it depends more on the degrees of freedom. The numerical results show that for different degrees of freedom the proposed test leads to the same conclusion, that is, the linear effect of age at diagnosis does not adequately capture its effect on the survival. In contrast, the test statistic of Arbogast and Lin's approach is 0.439 with the p-value of 0.512. This concludes that there is no strong evidence against the linear effect of age at diagnosis on the survival of the patients.

The scientific evidence (Adami et al., 1986) shows that the survival rates vary at different periods of observations and relative survival declines markedly as the women are getting older. However it is not clear whether the age at diagnosis plays a linear role in the logit link with survival status. Here we investigated this issue based on the cohort of 16,930 women. We selected three age groups $[40, 55)$, $[55, 70)$, and $[70, 90)$ on the entire cohort data of $16,930$ women, respectively. If age at diagnosis has a linear effect, we would expect the log odd-ratios to be not significantly different among these three groups. We fit the following linear logistic regression model to the data of each group

$$\begin{aligned} \text{pr}(Y = 1|X, Z, \text{Race}, \text{Grade}) &= H\{\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 I_d(\text{Race} = \text{white}) \\ &+ \sum_{k=1}^{3} \beta_{4k} I(\text{Grade} = k)\} . \end{aligned}$$

Here Y, X and Z are survival status, the binary variable corresponding to tumor marker 1, and age at diagnosis rescaled by dividing by 10, respectively. Here $I_d(\cdot)$ represents an indicator function.

The summary results of log odd-ratio parameter β_2 are listed in Table 11. Obviously the log odd-ratios for these three groups are quite different. Thus the data analysis from the cohort provides supplementary evidence in support of the conclusion from our score test that the effect of age at diagnosis on the survival status in the logistic regression model is

Table 11. Estimates from different age groups.

Age Group	Sample Size	Estimate of β_2	SE	p-value
$[40, 55)$	4598	0.106	0.096	0.268
$[55, 70)$	5396	0.643	0.077	< 0.001
$[70, 90)$	5806	1.068	0.056	< 0.001

strongly non-linear.

4.9 Discussion

In this chapter we proposed a test for testing adequacy of a functional form of a covariate in the logistic regression model of a matched case-control study. We applied the generalized score test method along with the regression spline technique used for approximating the nonparametric form of the covariate effect. The results of a simulation study indicate that the proposed method is more powerful than the existing method and computationally it is easy and less time consuming. As long as the number of knots is moderate, our proposed test is generally robust to the choice of knot points and the degrees of freedom. The analysis of the SEER breast cancer data not only illustrates the usefulness of the proposed method but also is scientifically important as we closely investigate the effect of age at diagnosis on the survival of the breast cancer patients which may shed new light on the pathogenesis of breast cancer.

Here we derived the form of the test in a simple set-up where the covariate of interest is observed without measurement error and has no missing values. In principle the proposed score method can be extended to the scenarios of partially missing data and when the covariate is measured with errors.

CHAPTER V

AN INFORMATION MATRIX BASED TEST IN MATCHED CASE-CONTROL STUDIES

5.1 Introduction

In Chapter IV we proposed a generalized score method to test the adequacy of a functional form of a covariate of interest in a conditional logistic regression model where the alternative model has the same link function as the null model. Thus we can write the null model as a special case of the alternative. When the link function of the alternative model is different from logistic function, for example, log link, probit, or inverse link, the generalized score method proposed in the previous chapter is not applicable anymore. Our goal in this chapter is to develop an information matrix based test for matched case-control studies.

For prospectively collected binary data, White (1982) proposed an information matrix test for detecting parametric model misspecification. Lin and Wei (1991) extended White's approach to the partial likelihood setting with particular interest in the Cox semiparametric regression model. Zhang (2001) applied this method to case-control studies. However, his method is not directly applicable to matched case-control studies.

In this chapter we will employ the idea to construct a test to check the validity of model

$$H_0 : \text{pr}(Y_{ij} = 1|\boldsymbol{S}_i, \boldsymbol{X}_{ij}) = H(\alpha_i(\boldsymbol{S}_i) + \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) \tag{5.1}$$

against the following alternative model in matched case-control studies

$$H_1 : \boldsymbol{P}(Y_{ij} = 1|\boldsymbol{S}_i, \boldsymbol{X}_{ij}) = f(\alpha_i(\boldsymbol{S}_i), \boldsymbol{X}_{ij}, \boldsymbol{\beta}), \tag{5.2}$$

where $f(\cdot)$ is an unknown function. For $j = 1, \cdots, M_i + 1, i = 1, \cdots, n$, Y_{ij} takes on

value one or zero according as the j^{th} subject in the i^{th} matched set with M_i controls is a case or control respectively, $\boldsymbol{X}_{ij} = (X_{ij1}, \cdots, X_{ijp})^{\mathrm{T}}$ is a $p \times 1$ vector of covariates, and $\boldsymbol{S}_i$ is the covariates which are used for matching purposes in the i^{th} stratum.

5.2 An Information Matrix Based Method

Under null model H_0, we again consider the conditional log likelihood function

$$L_{\mathrm{C}}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \sum_{j=1}^{M_i+1} p_{ij} Y_{ij}, \tag{5.3}$$

where $p_{ij} = \exp(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) / \sum_{k=1}^{M_i+1} \exp(\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta})$. Let $l(\boldsymbol{\beta})$ be the log of likelihood function $L_{\mathrm{C}}(\boldsymbol{\beta})$. Conditional on each stratum, i.e., given $\boldsymbol{X}$ and $\boldsymbol{S}$, define

$$\begin{aligned} \boldsymbol{B}(\boldsymbol{\beta}) &= -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{M_i+1} p_{ij} (\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{i\cdot})(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{i\cdot})^{\mathrm{T}} \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{V}(\boldsymbol{\beta}) &= \left\{ \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\} \left\{ \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \right\} \\ &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{M_i+1} (y_{ij} - p_{ij}) \boldsymbol{X}_{ij} \right\} \left\{ \sum_{j=1}^{M_i+1} (y_{ij} - p_{ij}) \boldsymbol{X}_{ij}^{\mathrm{T}} \right\} , \end{aligned}$$

where $\overline{\boldsymbol{X}}_{i\cdot} = \sum_{j=1}^{M_i+1} p_{ij} \boldsymbol{X}_{ij}$, and both $\boldsymbol{B}(\boldsymbol{\beta})$ and $\boldsymbol{V}(\boldsymbol{\beta})$ are symmetric. Since the conditional logistic regression model is a member of the exponential family of distributions, under Model (5.1), for the given strata $\boldsymbol{S}$ and covariate $\boldsymbol{X}$ we have

$$E\{\boldsymbol{B}(\boldsymbol{\beta})\} = E\{\boldsymbol{V}(\boldsymbol{\beta})\}\,.$$

To test the validity of null model, we can compare whether the two matrices are statistically different. Alternatively, we just need to compare the elements on or below the

diagonals of the two matrices. For $m_1 = 1, \cdots, n, m_2 \leq m_1$, let $b_{m_1 m_2}(\boldsymbol{\beta})$ and $v_{m_1 m_2}(\boldsymbol{\beta})$ be the components of matrix $\boldsymbol{B}(\boldsymbol{\beta})$ and $\boldsymbol{V}(\boldsymbol{\beta})$ at location of m_1^{th} row and m_2^{th} column, respectively. Then

$$b_{m_1 m_2}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{M_i+1} (X_{ijm_1} - \overline{X}_{i.m_1})(X_{ijm_2} - \overline{X}_{i.m_2})^{\mathrm{T}} p_{ij}$$

and

$$v_{m_1 m_2}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{M_i+1} (y_{ij} - p_{ij}) X_{ijm_1} \right\} \left\{ \sum_{j=1}^{M_i+1} (y_{ij} - p_{ij}) X_{ijm_2}^{\mathrm{T}} \right\},$$

where $\overline{X}_{i \cdot r} = \sum_{j=1}^{M_i+1} p_{ij} X_{ijr}$, i.e., the r^{th} component of $\overline{\boldsymbol{X}}_{i \cdot}$, $r = m_1, m_2$.

Notice that both $\overline{X}_{i \cdot r}$ and $\overline{\boldsymbol{X}}_{i \cdot}$ depend on $\boldsymbol{\beta}$, thus we have

$$\frac{\partial b_{m_1 m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{j=1}^{M_i+1} p_{ij} (X_{ijm_1} - \overline{X}_{i \cdot m_1})(X_{ijm_2} - \overline{X}_{i \cdot m_2})(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{i \cdot})$$

and

$$\begin{aligned} \frac{\partial v_{m_1 m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -\sum_{i=1}^{n} \left\{ \sum_{j=1}^{M_i+1} X_{ijm_1} p_{ij} (\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{i \cdot}) \right\} \left\{ \sum_{j=1}^{M_i+1} (Y_{ij} - p_{ij}) X_{ijm_2} \right\} \\ &\quad -\sum_{i=1}^{n} \left\{ \sum_{j=1}^{M_i+1} X_{ijm_2} p_{ij} (\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{i \cdot}) \right\} \left\{ \sum_{j=1}^{M_i+1} (Y_{ij} - p_{ij}) X_{ijm_1} \right\}, \end{aligned}$$

which implies

$$E\left\{ \partial v_{m_1 m_2}(\boldsymbol{\beta}) / \partial(\boldsymbol{\beta}) | \boldsymbol{X}_{ij} \right\} = 0$$

for $i = 1, 2, \ldots, n, j = 1, 2, \ldots, M_i + 1$.

Now Consider the lower triangular elements of matrix $\boldsymbol{V}(\boldsymbol{\beta}) - \boldsymbol{B}(\boldsymbol{\beta})$. Let

$$\begin{aligned} \boldsymbol{Q}_n(\boldsymbol{\beta}) &= (Q_{11}(\boldsymbol{\beta}),\, Q_{21}(\boldsymbol{\beta}),\, Q_{22}(\boldsymbol{\beta}),\, \cdots,\, Q_{pp}(\boldsymbol{\beta})\,) \\ &= (v_{11}(\boldsymbol{\beta}) - b_{11}(\boldsymbol{\beta}), v_{21}(\boldsymbol{\beta}) - b_{21}(\boldsymbol{\beta}), v_{22}(\boldsymbol{\beta}) - b_{22}(\boldsymbol{\beta}), \cdots, v_{pp}(\boldsymbol{\beta}) - b_{pp}(\boldsymbol{\beta}))\,. \end{aligned}$$

Suppose $\widehat{\boldsymbol{\beta}}$ is the maximum conditional likelihood estimator under Model (5.1) which is

obtained by maximizing (5.3). By Taylor expansion,

$$\mathbf{0} = \frac{\partial l(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial^T \boldsymbol{\beta}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \boldsymbol{O}_p(1) \ ,$$

we have

$$\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= -\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial^T \boldsymbol{\beta}} \right)^{-1} \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + o_p(n^{-1/2}) \\ &= \boldsymbol{B}^{-1}(\boldsymbol{\beta}) \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + o_p(n^{-1/2}) \end{aligned}$$

Thus

$$\begin{aligned} Q_{m_1 m_2}(\widehat{\boldsymbol{\beta}}) &= Q_{m_1 m_2}(\boldsymbol{\beta}) + \frac{\partial^{\mathrm{T}} Q_{m_1 m_2}}{\partial \boldsymbol{\beta}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(n^{-1/2}) \\ &= Q_{m_1 m_2}(\boldsymbol{\beta}) - \left(\frac{\partial b_{m_1 m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T \boldsymbol{B}^{-1}(\boldsymbol{\beta}) \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + o_p(n^{-1/2}) \end{aligned}$$

It can be shown that under Model (5.1), asymptotically

$$\boldsymbol{Q}_n(\widehat{\boldsymbol{\beta}})^T \widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\beta}}) \boldsymbol{Q}_n(\widehat{\boldsymbol{\beta}}) \to \chi^2_d \, ,$$

where $\Sigma(\boldsymbol{\beta}) = \{\sigma_{ij}\}$ is the $d \times d$ asymptotic covariance matrix of $\boldsymbol{Q}_n(\widehat{\boldsymbol{\beta}})$ that has a complicated form with $d = p(p+1)/2$ and can be estimated by $\widehat{\Sigma}(\widehat{\boldsymbol{\beta}})$. The derivation of Σ is given in Section 5.3.

5.3 Covariance Matrix Expression

Let $d(x)$ be the indicator function, $T(t)$ be the integer part of $t > 0$, then $i = \{m_1 - d(T(m_1/2) = m_1/2)\}T(m_1/2) + m_2$ and $j = \{r_1 - d(T(r_1/2) = r_1/2)\}T(r_1/2) + r_2$.

Since

$$
\begin{aligned}
Q_{m_1m_2}(\widehat{\boldsymbol{\beta}}) &= v_{m_1m_2}(\boldsymbol{\beta}) - b_{m_1m_2}(\boldsymbol{\beta}) - \left(\frac{\partial b_{m_1m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{B}^{-1}(\boldsymbol{\beta})\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\,, \\
Q_{r_1r_2}(\widehat{\boldsymbol{\beta}}) &= v_{r_1r_2}(\boldsymbol{\beta}) - b_{r_1r_2}(\boldsymbol{\beta}) - \left(\frac{\partial b_{r_1r_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{B}^{-1}(\boldsymbol{\beta})\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\,,
\end{aligned}
$$

we have

$$
\begin{aligned}
\sigma_{ij} &= \text{Cov}\left(Q_{m_1m_2}(\widehat{\boldsymbol{\beta}})\,,\, Q_{r_1r_2}(\widehat{\boldsymbol{\beta}})\right) \\
&= \text{Cov}\left\{v_{m_1m_2}(\boldsymbol{\beta}), v_{r_1r_2}(\boldsymbol{\beta})\right\} - \text{Cov}\left(v_{m_1m_2}(\boldsymbol{\beta}), \frac{\partial^T \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\boldsymbol{B}^{-1}(\boldsymbol{\beta})\frac{\partial b_{r_1r_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&\quad - \text{Cov}\left(v_{r_1r_2}(\boldsymbol{\beta}), \frac{\partial^T \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\boldsymbol{B}^{-1}(\boldsymbol{\beta})\frac{\partial b_{m_1m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&\quad + \left\{\frac{\partial b_{r_1r_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^T \boldsymbol{B}^{-1}(\boldsymbol{\beta})\text{var}\left\{\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}\left\{\frac{\partial b_{r_1r_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}\boldsymbol{B}^{-1}(\boldsymbol{\beta}) \\
&= E\{v_{m_1m_2}(\boldsymbol{\beta})v_{r_1r_2}(\boldsymbol{\beta})\} - E\left\{v_{m_1m_2}(\boldsymbol{\beta})\right\}E\left\{v_{r_1r_2}(\boldsymbol{\beta})\right\} - E\left\{v_{m_1m_2}(\boldsymbol{\beta})\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})\right\}^{\text{T}} \\
&\quad \times \boldsymbol{B}^{-1}(\boldsymbol{\beta})\frac{\partial b_{r_1r_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - E\left\{v_{r_1r_2}(\boldsymbol{\beta})\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})\right\}^T \boldsymbol{B}^{-1}(\boldsymbol{\beta})\frac{\partial b_{m_1m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&\quad + \left\{\frac{\partial b_{m_1m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^T \boldsymbol{B}^{-1}(\boldsymbol{\beta})\left\{\frac{\partial b_{m_1m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\} \\
&= E\left\{v_{m_1m_2}(\boldsymbol{\beta})v_{r_1r_2}(\boldsymbol{\beta})\right\} - b_{m_1m_2}(\boldsymbol{\beta})b_{r_1r_2}(\boldsymbol{\beta}) - E\left\{v_{m_1m_2}(\boldsymbol{\beta})\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^T \\
&\quad \times \boldsymbol{B}^{-1}(\boldsymbol{\beta})\frac{\partial b_{r_1r_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - E\left\{v_{r_1r_2}(\boldsymbol{\beta})\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})\right\}^T \boldsymbol{B}^{-1}(\boldsymbol{\beta})\frac{\partial b_{m_1m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&\quad + \left\{\frac{\partial b_{m_1m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^T \boldsymbol{B}^{-1}(\boldsymbol{\beta})\left\{\frac{\partial b_{m_1m_2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}.
\end{aligned}
$$

Here

$$
\begin{aligned}
& E\left\{v_{m_1m_2}(\boldsymbol{\beta})v_{r_1r_2}(\boldsymbol{\beta})\right\} \\
= & \sum_{i_1,i_2=1}^{n}\sum_{j,k,l,t=1}^{M_i+1} E\left\{(Y_{i_1j}-p_{i_1j})(Y_{i_1k}-p_{i_1k})(Y_{i_2l}-p_{i_2l})(Y_{i_2t}-p_{i_2t})\right\} \\
& \times \left\{X_{i_1jm_1}X_{i_1km_2}X_{i_2lr_1}X_{i_2tr_2}\right\} \\
= & \sum_{i=1}^{n}\sum_{j,k,l,t=1}^{M_i+} E\left\{(Y_{ij}-p_{ij})(Y_{ik}-p_{ik})(Y_{il}-p_{il})(Y_{it}-p_{it})\right\} X_{ijm_1}X_{ikm_2}X_{ilr_1}X_{itr_2} \\
& + \sum_{i_1\neq i_2}^{n} E\left\{\sum_{j,k=1}^{M_i+1}(Y_{i_1j}-p_{i_1j})(Y_{i_1k}-p_{i_1k})X_{i_1jm_1}X_{i_1km_2}\right\} \\
& \times E\left\{\sum_{j,k=1}^{M_i+1}(Y_{i_2j}-p_{i_2j})(Y_{i_2k}-p_{i_2k})X_{i_2jr_1}X_{i_2kr_2}\right\} \\
= & \sum_{i=1}^{n}\Delta_i(\boldsymbol{\beta}) + \sum_{(i_1\neq i_2)=1}^{n} K(i_1,\boldsymbol{\beta})K(i_2,\boldsymbol{\beta})\ ,
\end{aligned}
$$

where

$$
\begin{aligned}
\Delta_i(\boldsymbol{\beta}) \;=\; & \sum_{j,k,l,t=1}^{M_i+1} E\left\{(Y_{i_1j}-p_{i_1j})(Y_{i_1k}-p_{i_1k})(Y_{i_2l}-p_{i_2l})(Y_{i_2t}-p_{i_2t})\right\} \\
& \times X_{i_1jm_1}X_{i_1km_2}X_{i_2lr_1}X_{i_2tr_2}\ .
\end{aligned}
$$

Now we consider the $\Delta_i(\boldsymbol{\beta})$ in the following cases:

1. If $j=k=l=t$, then

$$
\delta_{i_1} = \sum_{j=1}^{M_i+1} p_{ij}(1-4p_{ij}+6p_{ij}^2-3p_{ij}^3)X_{ijm_1}X_{ijm_2}X_{ijr_1}X_{ijr_2}\ ;
$$

2. If only three of j,k,l,t are equal,

$$
\begin{aligned}
\delta_{i_2} \;=\; & -\sum_{j\neq k}^{M_i+1} p_{ij}p_{ik}(1-3p_{ij}+3p_{ij}^2)\{X_{ijm_1}X_{ijm_2}X_{ijr_1}X_{ikr_2} \\
& + X_{ijm_1}X_{ijm_2}X_{ikr_1}X_{ijr_2} + X_{ijm_1}X_{ikm_2}X_{ijr_1}X_{ijr_2} \\
& + X_{ikm_1}X_{ijm_2}X_{ijr_1}X_{ijr_2}\}\ ;
\end{aligned}
$$

3. If only two pairs of j, k, l, t are equal, for example $j = k, l = t, and \quad k \neq l$,

$$\begin{aligned}\delta_{i_3} &= -\sum_{j\neq k}^{M_i+1} p_{ij}p_{ik}(p_{ij}+p_{ik}-3p_{ij}p_{ik})\{X_{ijm_1}X_{ikm_2}X_{ijr_1}X_{ikr_2} \\ &\quad + X_{ijm_1}X_{ikm_2}X_{ikr_1}X_{ijr_2} + X_{ijm_1}X_{ijm_2}X_{ikr_1}X_{ikr_2}\}\,;\end{aligned}$$

4. If only two of j, k, l, t are equal, and $M_i \geq 2$,

$$\begin{aligned}\delta_{i_4} &= -\sum_{j\neq k\neq l}^{M_i+1} p_{ij}p_{ik}p_{il}(1-3p_{ij})\{X_{ijm_1}X_{ijm_2}X_{ilr_1}X_{ikr_2} + X_{ijm_1}X_{ikm_2} \\ &\quad \times X_{ijr_1}X_{ilr_2} + X_{ijm_1}X_{ikm_2}X_{ilr_1}X_{ijr_2} + X_{ikm_1}X_{ijm_2}X_{ijr_1}X_{ilr_2} \\ &\quad + X_{ikm_1}X_{ijm_2}X_{ilr_1}X_{ijr_2} + X_{ilm_1}X_{ikm_2}X_{ijr_1}X_{ijr_2}\}\,;\end{aligned}$$

5. If none of j, k, l, t are equal, and $M_i \geq 3$,

$$\delta_{i_5} = -3\sum_{j\neq k\neq l\neq t}^{M_i+1} p_{ij}p_{ik}p_{il}p_{it}X_{ijm_1}X_{ikm_2}X_{ilr_1}X_{itr_2}\,;$$

Thus $\Delta_i = \delta_{i_1} + \delta_{i_2} + \delta_{i_3} + \delta_{i_4} + \delta_{i_5}$, and

$$\begin{aligned}K(i,\boldsymbol{\beta}) &= E\left\{\sum_{j,k=1}^{M_i+1}(Y_{ij}-p_{ij})(Y_{ik}-p_{ik})\right\}X_{ijr_1}X_{ikr_2} \\ &= \sum_{j,k=1}^{M_i+1} E\left\{(Y_{ij}-p_{ij})(Y_{ik}-p_{ik})\right\}X_{ijr_1}X_{ikr_2} \\ &= \sum_{j=1}^{M_i+1} E\left\{(Y_{ij}-p_{ij})^2\right\}X_{ijr_1}X_{ijr_2} \\ &\quad + \sum_{j\neq k}^{M_i+1} E\left\{(Y_{ij}-p_{ij})(Y_{ik}-p_{ik})\right\}X_{ijr_1}X_{ikr_2} \\ &= \sum_{j=1}^{M_i+1} p_{ij}(1-p_{ij})X_{ijr_1}X_{ijr_2} - \sum_{j\neq k}^{M_i+1} p_{ij}p_{ik}X_{ijr_1}X_{ikr_2}\,.\end{aligned}$$

$$
\begin{aligned}
&\text{Cov}\left(v_{m_1m_2}(\boldsymbol{\beta}), \frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) = E\left\{v_{m_1m_2}(\boldsymbol{\beta})\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\} \\
= &\sum_{i_1,i_2=1}^{n}\sum_{j,k,l=1}^{M_i+1} E\left\{(Y_{i_1j}-p_{i_1j})(Y_{i_1k}-p_{i_1k})(Y_{i_2l}-p_{i_2l})\right\} X_{i_1jm_1}X_{i_1km_2}\boldsymbol{X}_{i_2l} \\
= &\sum_{i=1}^{n}\sum_{j,k,l=1}^{M_i+1} E\left\{(Y_{ij}-p_{ij})(Y_{ik}-p_{ik})(Y_{il}-p_{il})\right\} X_{ijm_1}X_{ikm_2}\boldsymbol{X}_{il} \\
= &\sum_{i=1}^{n}\boldsymbol{J}_i(\boldsymbol{\beta})\,.
\end{aligned}
$$

$\boldsymbol{J}_i(\boldsymbol{\beta})$ can be computed by the following cases:

1. If $j=k=l$, then

$$\gamma_{i_1} = \sum_{j=1}^{M_i+1} p_{ij}(1-3p_{ij}+2p_{ij}^2)X_{ijm_1}X_{ijm_2}\boldsymbol{X}_{ij}\,;$$

2. If only two of i,k,l are equal, then

$$\gamma_{i_2} = -\sum_{j\neq k}^{M_i+1} p_{ij}p_{ik}(1-2p_{ij})\{X_{ijm_1}X_{ijm_2}\boldsymbol{X}_{ik}+X_{ijm_1}X_{ikm_2}\boldsymbol{X}_{ij}+X_{ikm_1}X_{ijm_2}\boldsymbol{X}_{ij}\}\,;$$

3. If $i\neq k\neq l$, then

$$\gamma_{i_3} = 2\sum_{j\neq k\neq l}^{M_i+1} p_{ij}p_{ik}p_{il}X_{ijm_1}X_{ikm_2}\boldsymbol{X}_{il}\,,$$

So

$$\text{Cov}\left(v_{m_1m_2}(\boldsymbol{\beta}), \frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) = E\left\{v_{m_1m_2}(\boldsymbol{\beta})\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\} = \sum_{i=1}^{n}(\gamma_{i_1}+\gamma_{i_2}+\gamma_{i_3})\,.$$

5.4 A Simulation Study

In order to judge the performance of the methods under consideration we performed a simulation study by following the same scenarios in Section 4.7, Chapter IV.

We generated $N = 1000$ datasets for simulation purpose and compared the information matrix method IM with the cumulative residual based overall model adequacy test G_0 proposed by Arbogast and Lin (2004). Since the alternative models also has logit form, it would be interesting to include generalized score test T_n discussed in Chapter IV.

In scenario 1, we used only X to fit model to estimate the level and data were fit omitting X^2 for power. In scenario 2, we used X, Z to fit model to estimate the level and data were fit omitting X^2 for power. Since the limit distribution of G_0 follows a Gaussian process, for each dataset, we calculated p-value of the test based on 1000 bootstrap samples. For the calculation of the test statistic T_n and the degrees of freedom ν, one needs a value of the penalty parameter η. Following Gray's (1994) suggestion, we took 2.5 degrees of freedom and computed iteratively to obtain η. Also we used cubic splines and took 30% and 70% quantile as knots. The following are the simulation results. Table 12 corresponds to scenario 1 and Table 13 to scenario 2.

Table 12. Results of the simulation study for scenario 1 and $M = 3$.

β	N	TS	IM	G_0
0	25	0.053	0.048	0.047
0.10	25	0.161	0.071	0.062
0.25	25	0.562	0.252	0.210
0	50	0.051	0.062	0.062
0.10	50	0.224	0.115	0.097
0.25	50	0.852	0.457	0.464
0	100	0.063	0.052	0.051
0.10	100	0.345	0.163	0.167
0.25	100	0.991	0.764	0.790

The simulation results can be summarized as follows:

- These results indicate that the empirical levels of all three methods are not significantly different from the nominal level 0.05.
- The power of the tests is increasing with sample sizes as expected.

Table 13. Results of the simulation study for scenario 2 and $M = 3$.

β	N	TS	IM	G_0
0	25	0.058	0.046	0.049
0.10	25	0.149	0.073	0.041
0.25	25	0.573	0.212	0.181
0	50	0.048	0.059	0.051
0.10	50	0.232	0.101	0.078
0.25	50	0.856	0.354	0.412
0	100	0.054	0.057	0.056
0.10	100	0.373	0.124	0.151
0.25	100	0.987	0.596	0.778

- The finite sample performance of the overall model adequacy test IM is comparable to the overall model test G_0 of Arbogast and Lin (2004) for small parameters.
- In terms of computational time the IM method is much faster than that of Arbogast and Lin (2004), but is equivalent to score test T_n.
- Information matrix based method can be used to test the functional form of a continuous covariate, but it is not as powerful as score test T_n.

As expected, the overall model test is not as powerful as the score test. We have tried various sample sizes and different settings for alternative models to check the overall model fitting and compared the power with the method proposed by Arbogast and Lin (2004). Even though the information matrix method IM is much faster than the test G_0, it does not have great advantage in terms of power compared to cumulative residual based test G_0. Furthermore, information matrix based test can also test link functions, too. However, it may not be efficient in terms of power to test link functions. So a method of testing link function needs to be developed. Apart from the overall model test, how to handle the intercept term of each stratum in matched case-control is a main issue.

Overall, information matrix based test is suitable to explore the validity of overall

model fitting without taking too long time. If one need a more powerful test, one choice is to use cumulative residual based test G_0, otherwise a new method has to be developed.

CHAPTER VI

SUMMARY AND FUTURE RESEARCH

Case-control studies are widely used in epidemiological studies. In this dissertation we have considered the problems of bias reduction and goodness-of-fit measures.

The problem of bias reduction is presented in Chapter III. The purpose of this work is to reduce the bias of log-odds-ratio estimators in logistic regression model based on matched case-control studies. Usually, maximum conditional likelihood is used to estimate the log-odds-ratio. Due to the restriction of rare occurrence of disease or high cost of measurements, small to moderate sample sizes are not uncommon, which may cause significant bias problems.

We employ Firth's (1993) idea to matched case-control studies by penalizing the conditional likelihood with Jeffrey's invariant prior, and obtain the first order bias reduced estimator by solving the modified score function derived from the penalized conditional likelihood. The advantages of the proposed method are shown through a simulation study. We also apply the method to analyze a matched case-control data from a low-birth-weight study.

Chapter IV is developed to test the adequacy of a functional form of a covariate of interest in matched case-control studies. In matched case-control studies, we usually assume that the covariates are associated with the disease risk through a linear-logistic model, which means the logit of the disease probability is a linear function of the covariates. However, it is often not adequate to explain the effect of a continuous covariate on the disease risk. We develop a penalized score test for testing functional form of a covariate via a penalized regression spline with a moderately large number of given knot points. We study the asymptotic properties of the test and finite sample performance via a simulation study. It turns out that the power of the our proposed method is much better than that of the only one

existing method (Arbogast and Lin, 2004). We also illustrate the usefulness of the proposed method by using the SEER data. Furthermore, we develop a generalized method in Chapter V for testing overall goodness-of-fit of the logistic model for matched case-control studies. Here we use an information matrix based test by following the idea of White (1982) that dealt with the effect of model misspecification on maximum likelihood estimators. The finite sample performance of this test is judged via a simulation study. Importantly, both of my methods are computationally much faster than the existing ones.

In all the projects we assume that the data has complete observations. However, missing data or data with measurement errors are commonly seen almost in all research. Lipsitz et al. (1998) proposed a modified conditional logistic regression that is appropriate with covariates that are missing at random when the model involves many nuisance parameters. Paik and Sacco (2000) considered methods for analyzing matched case-control data when some covariates are completely observed but other covariates are missing for some subjects. Our proposed methods can be extended to handle bias and test model fitting issues that may appear in these scenarios.

Although our proposed methods are investigated in the context of matched case-control studies, the methods also are applicable in general case-control studies. In logistic regression model when covariates are measured with error the usual estimator is asymptotically biased (Michalik and Tripathi, 1980), Stefanski and Carroll (1985) proposed a bias-adjusted estimator to handle bias. By using the similar techniques to deal with the likelihood, it is possible to modify our proposed method to fit the model to get an estimator with less bias.

The proposed methods are also useful in analyzing the data from nested case-control studies where a case-control study is nested within a cohort study. A nested case-control study generally costs less and saves time compared to a cohort study. It also reduces recall bias compared with case-control studies.

In matched case-control study, we have considered how to test adequacy of a function form of covariates and overall model fitting. Even thought information matrix based can be used to test link function, it is not as powerful as the method in Arbogast and Lin (2004). In the future I am planning to develop a new method to test the link function.

Another problem of interest to me is family based case-control studies which are useful for studying familial aggregation of a disease. People have done a lots of work in this area (Whittemore, 1995; Zhao et al., 1998). However, still there are many questions on how to evaluate the effect of covariates on the disease risk, taking into account familial correlation using a flexible correlation structure and a logistic marginal model. And this problem has very practical applications in Epidemiology. For example, in Framingham Heart Studies, there are three generations involved. The causes of heart disease from the first or second generation may have an impact to the next generation. Even within the same generation, the members that are family related may show similar symptoms. Because of the correlation among family members, how to model the correlation within the model is critical for identifying the common factors or characteristics that contribute to the heart disease.

In the foreseeable future I plan to spend more time investigating case-control studies from clinical trials and Framingham Heart Studies. It is my strong desire to conduct cutting edge methodological and collaborative research.

REFERENCES

Adami, H. O., Malker, B., Holmberg, L., Persson, I., and Stone, B. (1986). The relation between survival and age at diagnosis in breast cancer. *The New England Journal of Medicine* **315**, 559–563.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1–10.

Arbogast, P. G. and Lin, D. Y. (2004). Goodness-of-fit methods for matched case-control studies. *The Canadian Journal of Statistics* **32**, 373–386.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* **53**, 370–418.

Bedrick, E. J. and Hill, J. R. (1996). Assessing the fit of the logistic regression model to individual matched sets of case-control data. *Biometrics* **52**, 1–9.

Bedrosian, I., Kuerer, H. M., Singletary, S. E., Hunt, K. K., Hortobagyi, G. N., and Babiera, G. (2008). Mammography before diagnosis among women age 80 years and older with breast cancer. *Journal of Clinical Oncology* **26**, 2482–2488.

Bishop, Y. M. M. adn Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analyses: Theory and Practice*. Cambridge, MA: MIT Press.

Boos, D. D. (1992). On generalized score tests. *The American Statistician* **46**, 327–333.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research*. Lyon, France: International Agency for Research on Cancer.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Pretencie, R. L., Sabai, C., and Aramesh, B. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology* **108**, 299–307.

Bull, S. B., Lewinger, J. P., and Lee, S. S. F. (2007). Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine* **26**, 903–918.

Bull, S. B., Mak, C., and Greenwood, C. M. T. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics & Data Analysis* **39**, 57–74.

Cordeiro, G. M. and McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society, Series B* **53**, 629–643.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

Cox, D. R. and Snell, E. J. (1968). *Analysis of Binary Data*. London: Chapman and Hall.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker Inc.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Goodwin, P. J., Ennis, M., Pritchard, K. I., Koo, J., Trudeau, M. E., and Hood, N. (2003). Diet and breast cancer, the evidence that extremes in diet are associated with poor survival. *Journal of Clinical Oncology* **21**, 2500–2507.

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics* **50**, 640–652.

Greenland, S. (2000). Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators. *Biostatistics* **1**, 113–122.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine* **25**, 4216–4226.

Heinze, G. and Schemper, M. (2001a). A solution to the problem of monotone likelihood in cox regression. *Biometrics* **57**, 114–119.

Heinze, G. and Schemper, M. (2001b). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.

Hosmer, D. W. and Lemeshow, S. (1985). Goodness-of-fit tests for the logistic regression model for matched case-control studies. *Biometrical Journal* **27**, 511–520.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Jewell, N. P. (1984). Small-sample bias of point estimators of the odds ratio from matched sets. *Biometrics* **40**, 421–435.

Kim, I. Y., Cohen, N. D., and Carroll, R. J. (2003). Semiparametric regression splines in matched case-control studies. *Biometrics* **59**, 1158–1169.

Liddell, F. D. K., McDonald, J. C., and Thomas, D. C. (1977). Methods of cohort analysis: appraisal by application to asbestos mining. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **140**, 469–491.

Lin, D. Y. and Wei, L. J. (1991). Goodness-of-fit for the general cox regression model. *Statista Sinica* **1**, 1–17.

Lipsitz, S. R., Parzen, M., and Ewell, M. (1998). Inference using conditional logistic regression with missing covariates. *Biometrics* **54**, 295–303.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157.

Michalik, J. E. and Tripathi, R. C. (1980). The effect of errors in gianosis and measurement on the estimation of the probability of an event. *Journal of the American Statistical Association* **75**, 713–721.

Miettinen, O. (1970). Estaimation of relative risk from individually matched series. *Biometrics* **26**, 75–86.

Moolgavkar, S. H., Lustbade, E. D., and Venzon, D. J. (1984). A geometric approach to nonlinear regression diagnostics with application to matched case-control studies. *The Annals of Statistics* **12**, 816–826.

Moolgavkar, S. H., Lustbade, E. D., and Venzon, D. J. (1985). Assessing the adequacy of the logistic regression model for matched case-control studies. *Statistics in Medicine* **4**, 425–435.

Paik, M. C. and Sacco, R. L. (2000). Matched case-control data analyses with missing covariate. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **49**, 145–156.

Pike, M. and Morrow, R. H. (1970). Statistical analysis off patient-control studies in epidemiology. *British Journal of Preventive & Social Medicine* **24**, 42–44.

Pregibon, D. (1985). Data analysis methods for matched case-control studies. *Biometrics* **40**, 639–651.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons, Inc.

Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics* **13**, 1335–1351.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **86**, 531–539.

Whittemore, A. S. (1995). Logistic regression of family data from case-control studies. *Biometrika* **82**, 57–67.

Woolf, B. (1955). On estimating the relationship between blood group and disease. *Annals of Human Genetics* **16**, 251–253.

Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society* **1**, 217–235.

Zhang, B. (2001). An information matrix test for logistic regression models based on case-control data. *Biometrika* **88**, 921–932.

Zhao, L. P., Hsu, L., Holte, S., and Chen, Y. (1998). Combined association and aggregation analysis of data from case-control family studies. *Biometrika* **85**, 299–315.

VITA

Xiuzhen Sun was born in Shandong Province, China. After she finished her B.S. degree in Mathematics at Shandong Normal University, China, 1991, she taught mathematics to undergraduate students at Petroleum University, China. She came to the United States in January 2002. In the spring of 2003, she entered the graduate program in mathematics at Southern Methodist University at Dallas, Texas, and graduated with a Master of Science degree in Applied Mathematics in May 2005. She came to the Department of Statistics, Texas A&M University in College Station, Texas in August 2005 and received a Master of Science degree in Statistics in May 2007. She then pursued her Doctor of Philosophy degree completed in August 2010 under the supervision of Dr. Suojin Wang and Dr. Samiran Sinha. In August, she began work at the Biostatistics Department, Boston University School of Public Health for two years as a postdoctoral fellow. Her mailing address is:

Biostatistics Department
Boston University School of Public Health
801 Mass Avenue, Crosstown, 3^{th} floor
Boston, MA 02118